OFFICE OF PEOPLE ANALYTICS

# Next Generation Testing: Overview and Update

Mary Pommerich & Tia Fechter
*Defense Personnel Assessment Center*

Scott Oppler
*HumRRO*

DAC Meeting
September 18, 2020

# Purpose and Overview

- Provide background and update on history, status, and plans for the next generation of ASVAB and special tests administered on the ASVAB platform† in the military's Enlistment Testing Program (ETP).
  - Discuss changes since the 2005–2006 ASVAB review and status of current review efforts (Mary Pommerich).
  - Discuss *Next Generation Testing* efforts.
    - Give status report on the evaluation of ASVAB tests (Mary Pommerich).
    - Discuss thoughts on consolidating ASVAB evaluation findings into one rating (Tia Fechter).
    - Discuss focus group effort to develop a shared vision for *Next Generation Testing* across stakeholders (Scott Oppler).

†ASVAB platform = The test delivery modality for the ASVAB and various special tests administered in the military's Enlistment Testing Program (ETP).

# ASVAB Review History

- The ASVAB underwent a systematic review in 2005–2006, with testing experts making recommendations for improvements and enhancements to the military's Enlistment Testing Program (ETP).

*The panel was motivated by a difficult recruiting environment and the belief held by some that the ASVAB was outdated and in need of an overhaul.*

- The Manpower Accession Policy Working Group (MAPWG) condensed and prioritized the Panel's recommendations.
  - A modified Delphi approach* was used to prioritize the condensed recommendations.

*The Delphi approach will be discussed later in the briefing.

# MAPWG Prioritized Recommendations and Status

- Implement CAT at MET sites [1]
- Consider classification accuracy when evaluating content changes [1]
- Re-evaluate the contents of the ASVAB [1]
- Examine validity regularly [5]
- Increase time for seeding new items and measures [5]
- Include validated non-cognitive measures in job classification composites [7]
- Include nonverbal reasoning test on ASVAB [8]
- Develop standardized data banks on Service member performance [8]
- Relax the requirement for criterion validity of new measures [8]
- Implement content controls in CAT [8]
- Continue utility research on non-cognitive measures [12]
- Develop IT/communications technology test [13]
- Review test specifications on a regular basis [13]
- Evaluate WK and PC for ESL examinees [13]
- Consider the multidimensionality of the ASVAB [13]
- Evaluate Spanish verbal test for ESL examinees [17]
- Use automatic item generation [17]

Recommendation has been/is being/will be implemented at an appropriate time.

Development of a nonverbal reasoning test (Complex Reasoning) is underway.

A congressionally mandated effort is underway to review the applicability of current military testing practices to the English language learner population.

# MAPWG Prioritized Recommendations and Status

- Many changes have been introduced in the Enlistment Testing Program as a result of the 2005–2006 review.
- The contents of the *ASVAB itself* have not yet changed. Prior discussions of possible changes have floundered on:
  - Lack of consensus on the philosophy of the ASVAB.
  - Logistical difficulties associated with making changes (such as dropping subtests) that would impact existing composites and systems set up to operate on those composites.
  - Concerns about insufficient resources to accommodate a revised ASVAB that would take more time than the current battery (if new tests of interest were added to the current battery).
- Given the complexities associated with making changes to the ASVAB, DPAC now believes it is best to consider all new and existing tests at once, rather than on a case-by-case basis.

# Next Generation Testing Effort

- DPAC hopes to resolve the ASVAB impasse via *Next Generation Testing* efforts.
- Key steps:
  1. Study new tests of interest.
  2. Evaluate the tests in the ASVAB.
  3. Consolidate information gathered to aid decision-making about the status of individual tests.
  4. Conduct focus groups with stakeholders to develop a shared vision for *Next Generation Testing*.

# Next Generation Testing Questions

What tests should be administered as part of the ASVAB or on the ASVAB platform in the future?

What other changes are needed to modernize the ASVAB/ETP?

# IT'S NOT *JUST* ABOUT THE ASVAB

- *Next Generation Testing* efforts will focus on the *ASVAB*, as well as the *special tests* that are administered alongside the ASVAB in the ETP.
  - ASVAB and special tests are administered jointly on the ASVAB platform and share a common look.
  - Special test scores are used in addition to ASVAB scores for classification purposes.
  - A key distinction is who is responsible for development and maintenance of the tests.
    - DPAC has responsibility for the ASVAB; Service proponents have responsibility for the special tests.
- Due to the limited time for total testing, it is necessary to consider all tests to be administered on the ASVAB platform in conjunction.

# Next Generation Testing – Progress Report

- Continue efforts to develop or refine new tests of interest (TAPAS, Cyber Test, Mental Counters, Complex Reasoning).

  ☑ *Ongoing.*

- Continue efforts to evaluate tests currently in the ASVAB.

  ☑ *Ongoing. Details to follow.*

- Complete effort to apply an argument-based approach to validation of the ASVAB.

  ☑ *Ongoing. Has been completed for AFQT tests.*

- Review and update the psychometric checklist, as needed, for the purpose of evaluating tests to be administered as part of the ASVAB.

  ☑ *Ongoing: Updates and revisions incorporated in 2019 and 2020.*

# Next Generation Testing – Progress Report

- Services/proponents complete the updated psychometric checklist for new tests of interest, documenting all new information since a checklist was previously completed.

  🗁 *Initial checklists available for updating by proponents.**

- Stakeholders develop a shared vision that defines the purpose and general makeup of the ASVAB and ETP for *Next Generation Testing.*

  🗁 *Initiated. Results of a MAPWG focus group and plans for future focus groups to be summarized later.*

- Establish a systematic process for evaluating potential changes and making decisions regarding tests in the ASVAB and the ETP.

  🗁 *Initiated. Details to follow.*

*Updates would ideally be done following completion of steps to address any concerns (e.g., see Slide 13).

# Next Generation Testing – Progress Report

- Revisit logistical questions with stakeholders, including the feasibility of lengthening the ASVAB and the feasibility of dropping existing tests.

  🕓 *Future effort to occur after ASVAB evaluation and focus groups.*

- Stakeholders summarize the impact of potential modifications to the battery and identify resources to support a revised battery.

  🕓 *Future effort to occur after ASVAB evaluation, evaluations of new tests of interest, and focus groups.*

- Compile all information, then identify, discuss, and move forward with potential changes to the contents of the ASVAB and the special tests administered in the ETP.

  🕓 *Future effort to occur after completion of all above steps.*

# New Tests of Interest

# New Tests of Interest – Recap

- The Services/DPAC are continuing to develop and/or fine-tune key new tests of interest:

**Cyber Test**

Updated forms are being introduced to address compromise and obsolescence concerns.

A CAT version will be introduced in the cloud to better target item difficulty to applicant ability.

**Mental Counters**

Improvements to the instructions and practice items are being studied to eliminate a persistent floor effect in the applicant population.

**Complex Reasoning**

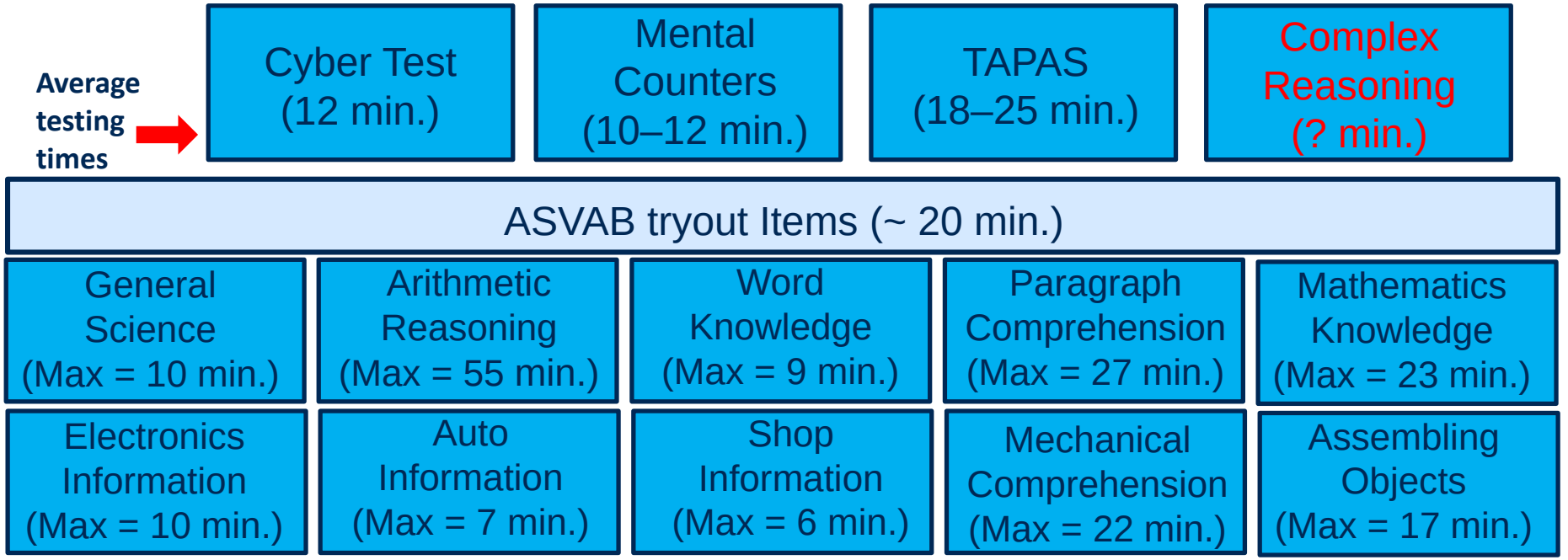
A non-verbal test of fluid intelligence is under development with items modeled after Raven's Progressive Matrices items.

**TAPAS**

A way forward is being discussed in response to recommendations from the TAPAS Evaluation Project.

# Testing Time Considerations

- Due to resource constraints, total testing time across the ASVAB and special tests (as well as potentially dated content) will be a key consideration for *Next Generation Testing*:




- Hence, there is a strong interest in assessing how the ASVAB might be modified to accommodate new tests.

# ASVAB Evaluation

# ASVAB Evaluation Motivation

- Research has been ongoing to thoroughly evaluate the new tests of interest, but the existing ASVAB tests have not systematically undergone similar scrutiny.

> A comprehensive assessment of the tests currently in the battery will give insight into their utility, quality, and potential modifiability.

- Potential changes to ASVAB to accommodate new tests for *Next Generation Testing* could include any combination of the following:

- Dropping existing tests
- Combining existing tests
- Shortening existing tests
- Merging new tests with existing tests

# ASVAB Contents – Evaluation Effort

- DPAC has an extensive effort underway to evaluate the current ASVAB tests in order to determine their desirability/expendability, including:
  - ❑ Reviewing the history of current ASVAB tests and why they were originally included in the battery.
  - ❑ Completing psychometric checklists and evaluating the psychometric value/limitations for each test.
  - ❑ Evaluating the usefulness/appropriateness of existing tests with the current applicant population.
  - ❑ Evaluating the item/form development costs.
  - ❑ Evaluating the ease/difficulty of developing good quality items.
  - ❑ Evaluating the durability of test content.
  - ❑ Evaluating the appropriateness/efficiency of content coverage across tests.
  - ❑ Evaluating the vulnerability of content to compromise and other unwanted effects.
  - ❑ Evaluating the efficiency of each test.
  - ❑ Evaluating the psychometric impact of shortening or combining various tests.
  - ❑ Evaluating the psychometric impact of dropping various tests.

# Project Team/Contributors

- Dan Segall (DPAC)
- Furong Gao (HumRRO)
- Greg Manley (DPAC)
- Jeff Harber (DPAC)
- Lihua Yao (DPAC)
- Mary Pommerich (DPAC)
- Matt Trippe (HumRRO)
- Ping Yin (HumRRO)
- Rich Riemer (DPAC)
- Robert Hamilton (DPAC)
- Sachi Phillips (HumRRO)
- Scott Oppler (HumRRO)
- Tia Fechter (DPAC)
- Tom Waterbury (HumRRO)

# Step 1: Trace History of Current Tests

- Goal: Document where the ASVAB tests came from and why they were originally included in the battery.
- Team: Tia Fechter, Greg Manley
- Resources:
  - ASVAB Working Group's *History of the ASVAB 1974–1980*
  - Bayroff and Fuchs (1970)
  - Maier & Sims (1986)
  - Maier (1993)
  - Uhlaner and Bolanovich (1952)
- Status: Completed ☑

Information regarding the provenance of all the current ASVAB tests has been found.

# Step 1: Trace History of Current Tests

| Test | 1968*–1975 (P&P) | 1976**–1980 (P&P) | 1980–2002 (P&P) | 2002–current (P&P) | 1990–current (CAT) |
|---|---|---|---|---|---|
| Word Knowledge (WK) | x | x | x | x | x |
| Arithmetic Reasoning (AR) | x | x | x | x | x |
| Mechanical Comprehension (MC) | x | x | x | x | x |
| Shop Information (SI) | x | x | x | x | x |
| Automotive Information (AI) | x | x | x | x | x |
| Electronics Information (EI) | x | x | x | x | x |
| Mathematics Knowledge (MK) | | x | x | x | x |
| General Science (GS) | | x | x | x | x |
| Paragraph Comprehension (PC) | | | x | x | x |
| Assembling Objects (AO) | | | | x | x |

*Introduction of the ASVAB for use in the STP (CEP)

**Introduction of the ASVAB for joint-Service use in the ETP

# Step 1: Trace History of Current Tests

- Goal: Document where the ASVAB tests came from and why they were originally included in the battery.
  - *WK, AR, MC, AI, SI, and EI were all included in the joint-Service STP-ASVAB in 1968 because these tests were identified from the various classification tests used across the Services as "interchangeable," which was considered important to multiple Services.*
    - *Arithmetic reasoning was one of three content areas included in the first Armed Forces Qualification Test (AFQT) introduced in 1950 (along with vocabulary and spatial relations).*
    - *Automotive Information, Mechanical Aptitude, and Clerical Speed were among the individual aptitude tests developed by the Services between 1941–1949 as supplementary tests to use in classification decisions. Research and operational experience suggested these tests were needed.*
  - *When the content of the joint-Service ETP-ASVAB was being decided in the 1974 ASVAB Working Group meetings, Army and Navy expressed the need for a science measure (the documentation does not elaborate why). To address the need, Forms 5/6/7 introduced the GS test containing physical science content from the Navy's Electronic Technician Selection Test and biological content from the Army's science test.*

# Step 1: Trace History of Current Tests

- Goal: Document where the ASVAB tests came from and why they were originally included in the battery.
  - *Originally, MK was added to aid in classification of Service members into military occupational specialties (MOSs). Later, MK was also added as part of the AFQT as a replacement for the Numerical Operations test that was being removed from the AFQT due to its complications (e.g., sensitivity to score differences as a result of various format differences, coaching, practice effects, cheating). MK was found to be a better predictor for general trainability and resulted in more accurate scores.*
  - *PC was included to increase the literacy requirements in the AFQT, in response to findings that recruits had difficulty reading the instructional materials in their training courses.*
  - *It is a popularly held belief that AO was selected, in part, because it was one of the few ECAT tests that could be administered across both CAT and P&P platforms. In reality, AO was one of the most promising of the 9 ECAT tests, when considering findings across all analyses and evaluations.*

# Step 2: Complete Psychometric Checklists

- Goal: Complete the psychometric checklist for current ASVAB tests and Coding Speed (CS) and evaluate psychometric value/limitations of each test.
- Team: Tia Fechter, Greg Manley, Sachi Phillips, Tom Waterbury
- Status:
  - Final checklists have been completed for AO, AR, MK and PC. ☑
  - Draft checklists have been completed for GS, WK, MC, EI, and CS. ☑
  - Draft checklist is in progress for AS (AI/SI).

# Step 2: Complete Psychometric Checklists

- Interim Step: Identify pros and cons for each test and synthesize.
- Status: Initial pros and cons have been identified for GS, AR, WK, PC, MK, EI, AI/SI, MC, and AO. ☑

Next Step:
- Complete remaining checklists, fine-tune pros and cons, and synthesize results across tests.

# Step 2: Complete Psychometric Checklists

PROS for MK:

- Less vulnerable to practice effects
- Less potential for adverse impact* than other ASVAB tests due to language-free content
- Minimal adverse impact* for M/F and NHW/NHA†
- Small adverse impact* for NHW/HW and NHW/NHB
- Requires no sensitivity review due to language-free content
- A good candidate for automated item generation
- Predictive validity with training criteria
- Good potential for classification efficiency

CONS for MK:

- Vulnerable to compromise
- Multidimensionality concerns, with no discernable content specificity
- Requires identification of item enemies to avoid local dependency issues
- Possible platform/device effects related to presentation of mathematical symbols and graphics

- Pros and cons lists for the other tests can be found in the backup slides.

*As defined by effect size

†Small adverse impact was observed for NHW/NHA, but is labeled minimal because it is in the direction that favors the minority

# Step 2: Synthesize Pros and Cons

## Pros

| Test | Unique Domain | Alternate Measure of Verbal Ability | Useful for AFQT Compromise Detection | Requires little test time | Provides incremental validity | Predictive Validity with Training Criteria | Good potential for classification efficiency | Minimal adverse impact for M/F | Minimal adverse impact for NHW/HW | Minimal adverse impact for NHW/NHB | Minimal adverse impact for NHW/NHA | Small adverse impact for M/F | Small adverse impact for NHW/HW | Small adverse impact for NHW/NHB | Small adverse impact for NHW/NHA | Long history as an AFQT measure | Exhibits lower magnitude of adverse impact compared to other verbal measures | Nonverbal test | Less potential for adverse impact due to language-free content | Sensitivity review not required | Excellent candidate for AIG | Content is stable over time | Items inexpensive to produce | Good candidate for AIG | Possible candidate for AIG | High item retention rates | Not vulnerable to practice effects | Less vulnerable to practice effects than other psychomotor or spatial tests | Resistant to coaching, cheating, or compromise | Less critical to update pools frequently |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| GS | x* | x | x | x | x | x | | | | | | | | | | | | | | | | | | | | | | | |
| AR | | | | | x | x | | | | | x | x | x | | | x | | | | | x | x | | | | | | | |
| WK | | | | x | | x | | | | | | x | | | | | | | | | | | x | x† | | x | x | | |
| PC | | | | | x | | | | | | | x | | | | | x | | | | | | | | x** | | | x | |
| MK | | | | | | x | x | x | | | x | | x | x | | | | | x | x | | | | x | | | x†† | | |
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| EI | | | | x | | | x | | | | | | | | | | | | | | | | | | | | | x | |
| AS (AI/SI) | | | | | x | | | | | | | | | | | | | | | | | | | | | | | x | |
| MC | | | | | | | | | | | | | | | x | | | | | | | | | | | x | | x | |
| AO | x | | | | x | x | x | x | x | | x | | | | | | | x | x | x | | | | x | | | x | x | x |

*Partially unique
†For definition item type
**For some item types
††Less vulnerable than other ASVAB tests

# Step 2: Synthesize Pros and Cons

## Cons

| Test | Multidimensionality Concerns/Issues | Extra time required to address MD concerns | Moderate adverse impact for M/F | Moderate adverse impact for NHW/HW | Moderate impact for NHW/NHB | Moderate impact for NHW/NHA | Large to very large adverse impact for M/F | Large to very large adverse impact for NHW/HW | Large to very large impact for NHW/NHB | Large to very large impact for NHW/NHA | Vulnerable to compromise | Vulnerable to practice effects | Shows greater magnitude of adverse impact than other test with the same domain | Possible platform/device effects | Lengthy time limits required | Requires frequent item replacement | Domain is finite or relatively finite | Can be difficult for English learners | Limited to one item per passage in CAT | Past issues with ceiling effects | Sensitivity concerns related to content | Durability concerns related to content | Requires extensive enemy lists | Not a good candidate for AIG |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| GS | x | | x | x | x | x | | | | | | | | | | | | | | | | | | |
| AR | | | | | x | | | | x | | x | x | x | x | x | | | | | | | | | |
| WK | | | | x | | x | | | x | | x | | | | | x | x | x | | | | | | |
| PC | x | | | x | x | x | | | | | | | | | x | | | | x | x | x | x | | |
| MK | x | | | | | | | | | | x | | | x | | | | | | | | | x | |
| EI | | | x | x | | x | | | x | | | | | | | | x* | | | | | x | | x |
| AS (AI/SI) | x | x | | x | | | x | x | x | | | | | | | | x* | | | | | x†† | | x |
| MC | | | x | x | | | | | x | | | | | | | | x* | | | | | | x | x |
| AO | x | x | | | | | | | | | | | | x | | | | | | x | | | | |

# Step 3: Evaluate Usefulness, Appropriateness

- Goal: Evaluate the usefulness and appropriateness of existing tests with regard to the current population.
- Task 3a: Track trends in test scores over years 1984–2018.
  - Team: Tia Fechter, Robert Hamilton, Lihua Yao
- Task 3b: Evaluate what fraction of the population possesses the knowledge/skill assessed by the test.
  - Task 3b(i): Evaluate overlap between latent ability and score information for current testing population.
    - Team: Mary Pommerich, Ping Yin
  - Task 3b(ii): Use job task analysis ratings to evaluate the relevance of content contained in the science and technical tests to success in technical training.
    - Team: Tia Fechter, Scott Oppler, Dan Segall

# Step 3: Evaluate Usefulness, Appropriateness

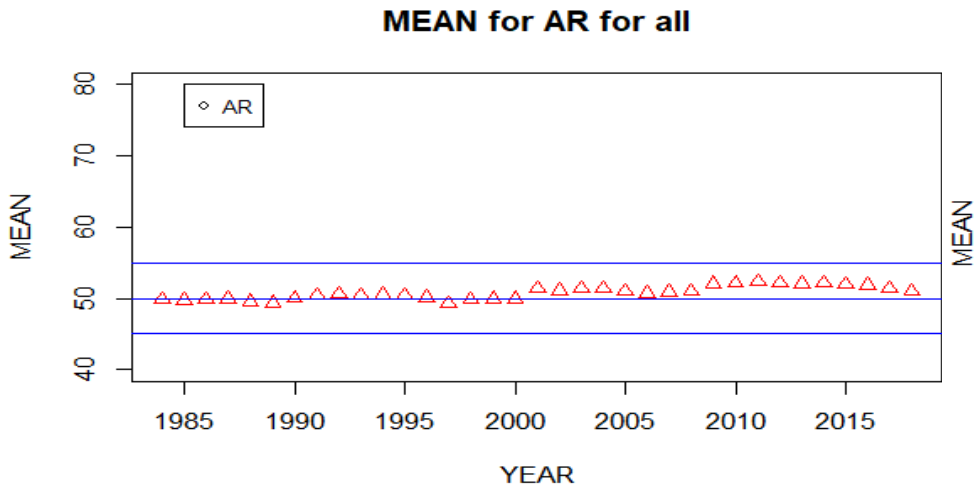
- Goal: Evaluate the usefulness and appropriateness of existing tests with regard to the current population.
- Task 3a: Track trends in test scores over years 1984–2018.
  - Status: Analyses completed ☑

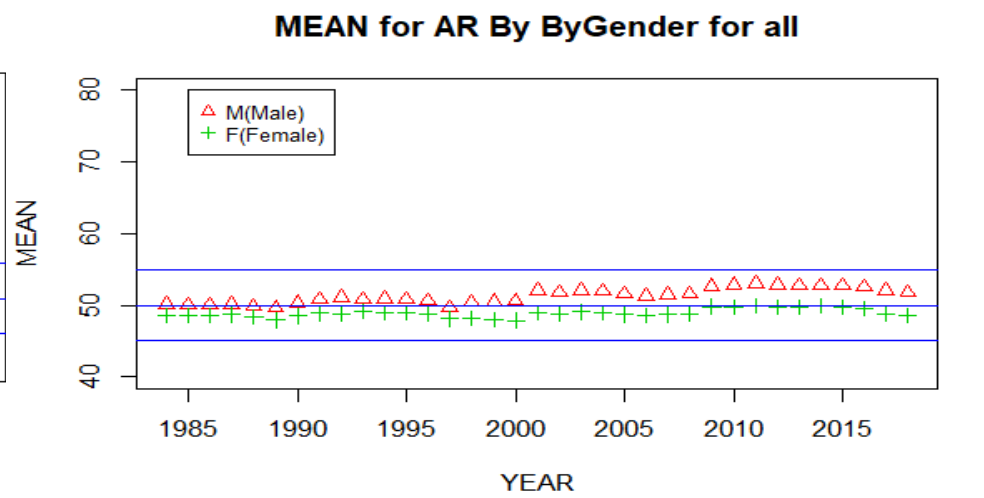
Located data:
- 1997–2018 for CAT-ASVAB, 1984–2018 for P&P-ASVAB (ETP only)

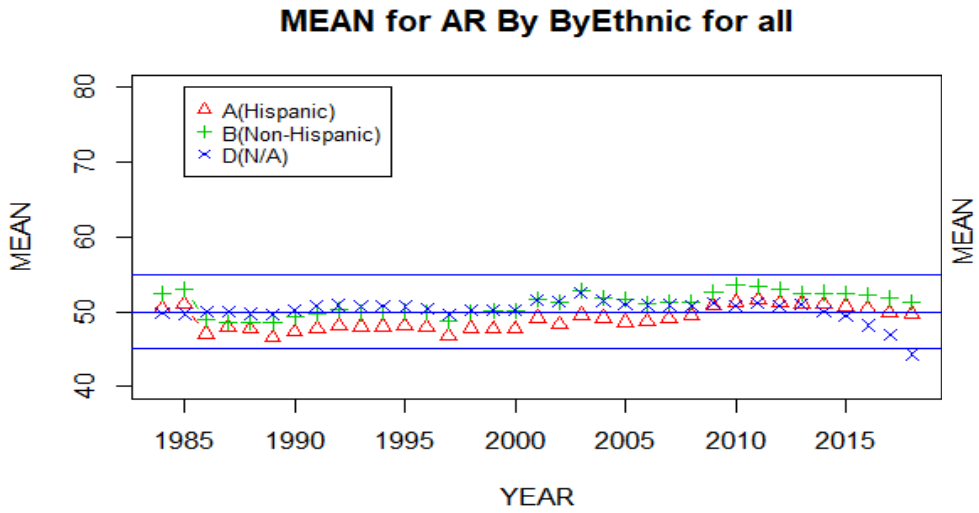
Scaled data:
- Scores on the PAY80 score scale were converted to the PAY97 scale

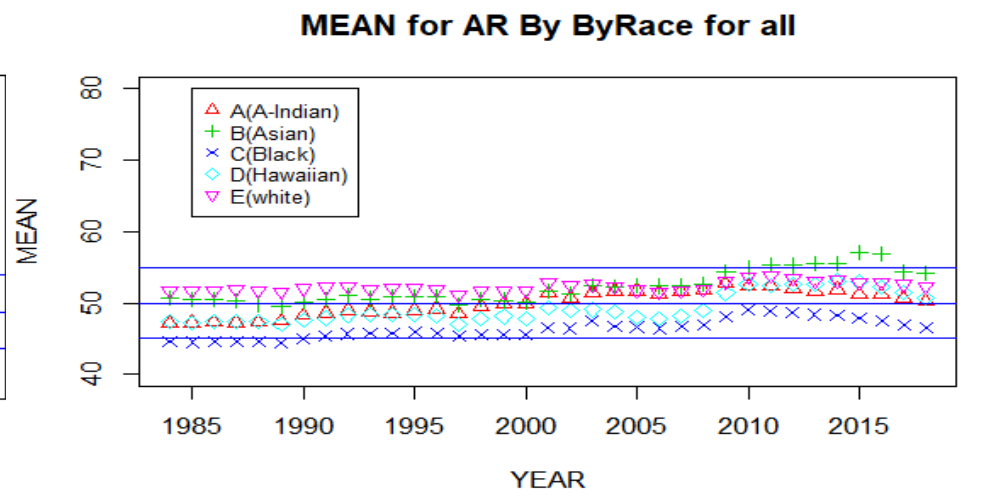
Analyzed data:
- Computed summary statistics for P&P-ASVAB, CAT-ASVAB, and combined by gender, ethnicity, and race
- Conducted ANOVA with AFQT as dependent variable and year as independent variable
  - Year and economy (unemployment rate) have a significant impact on AFQT scores

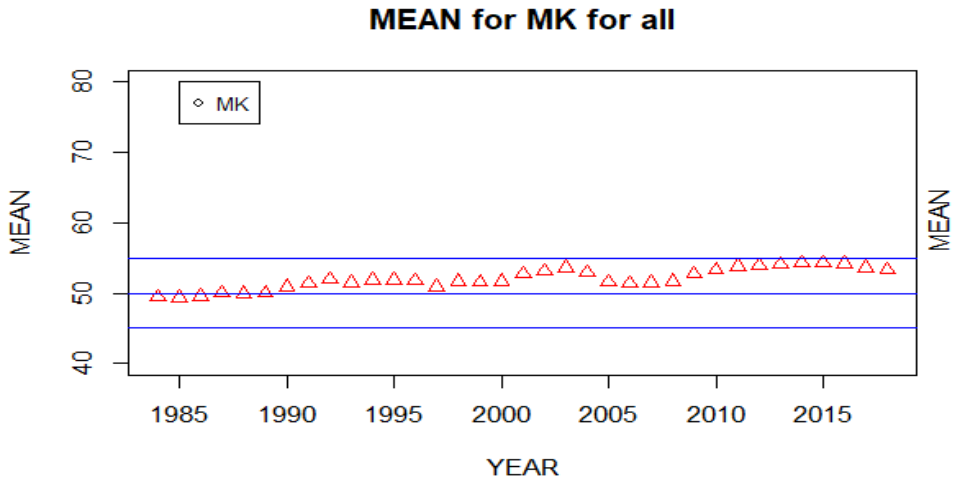
Plotted and summarized results for all tests

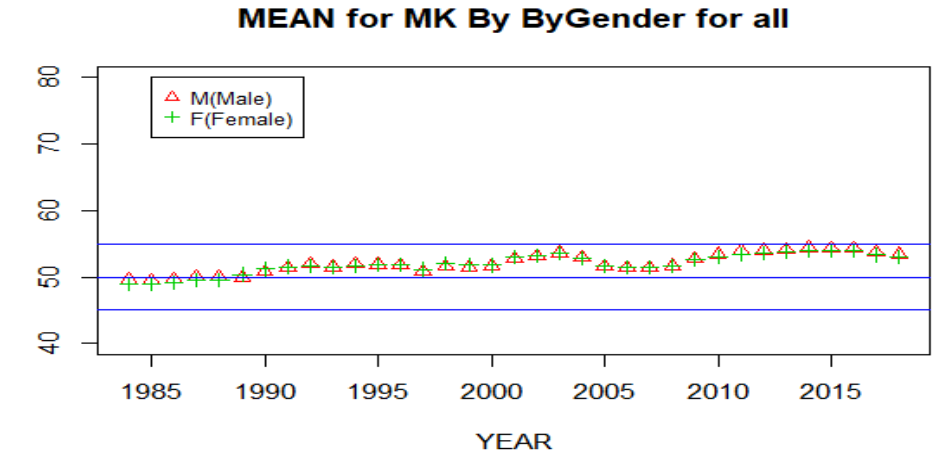

MEAN for AR for all
AR
MEAN
YEAR
40
50
60
70
80
1985
1990
1995
2000
2005
2010
2015

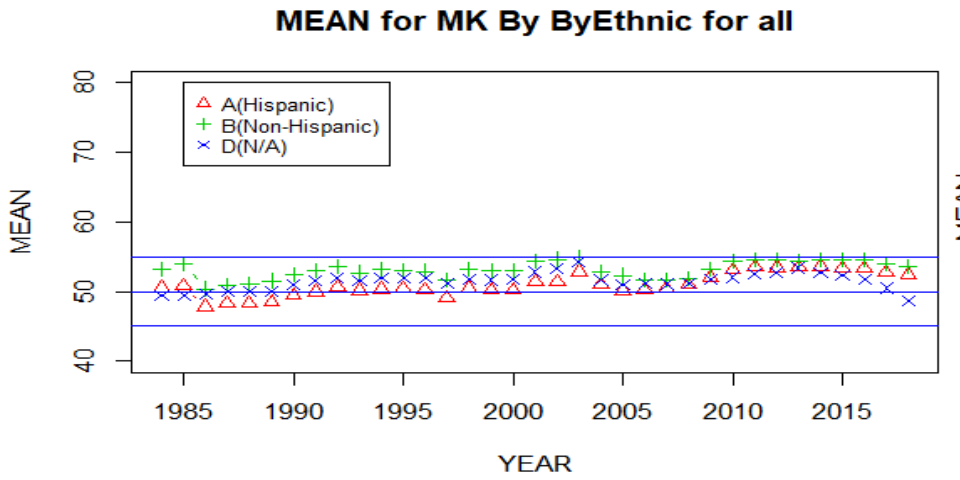

MEAN for AR By ByGender for all
M(Male)
F(Female)
MEAN
YEAR
40
50
60
70
80
1985
1990
1995
2000
2005
2010
2015

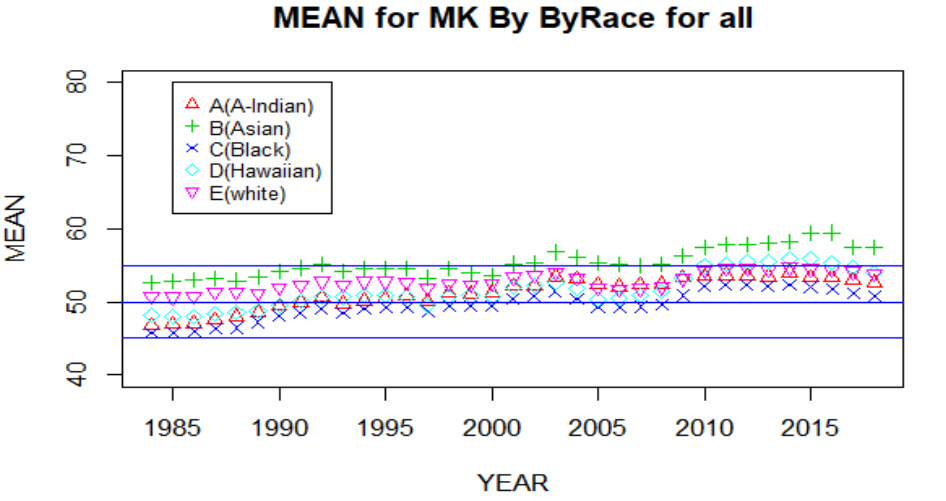

MEAN for AR By ByEthnic for all
A(Hispanic)
B(Non-Hispanic)
D(N/A)
MEAN
YEAR
40
50
60
70
80
1985
1990
1995
2000
2005
2010
2015

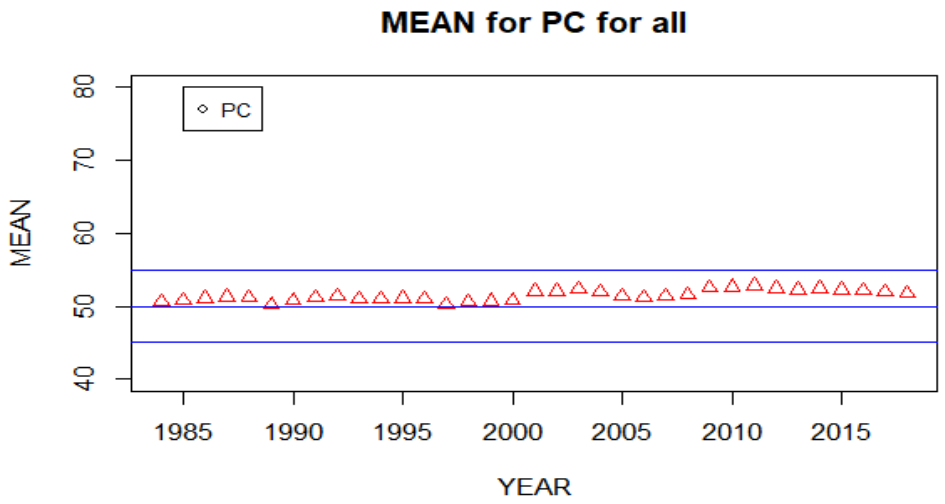

MEAN for AR By ByRace for all
A(A-Indian)
B(Asian)
C(Black)
D(Hawaiian)
E(white)
MEAN
YEAR
40
50
60
70
80
1985
1990
1995
2000
2005
2010
2015

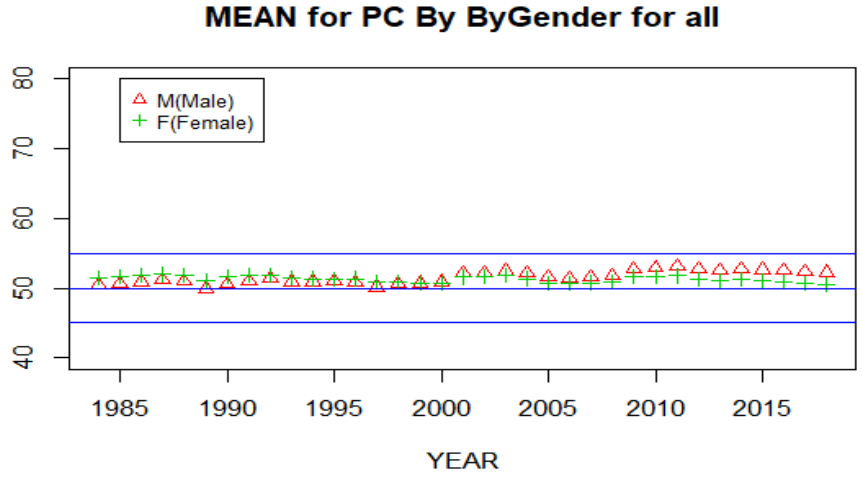

MEAN for MK for all
MK
MEAN
YEAR
40
50
60
70
80
1985
1990
1995
2000
2005
2010
2015

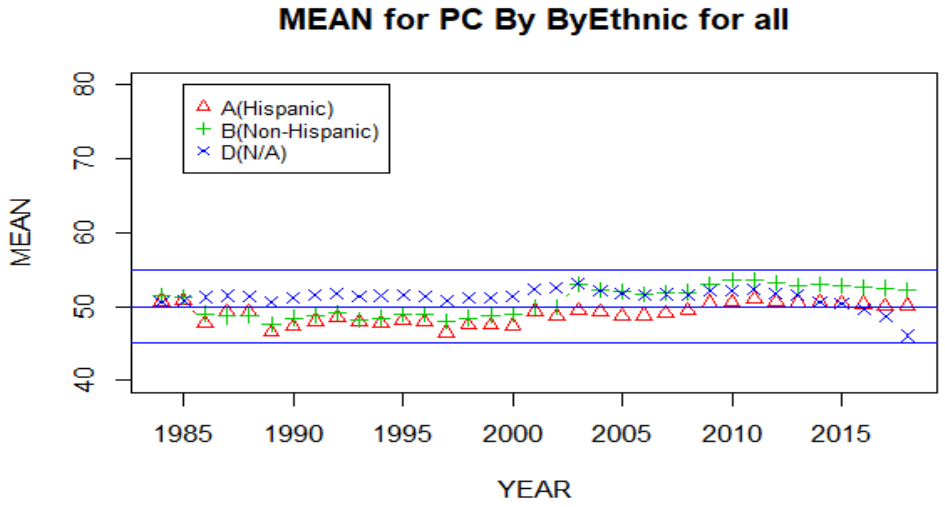

MEAN for MK By ByGender for all
M(Male)
F(Female)
MEAN
YEAR
40
50
60
70
80
1985
1990
1995
2000
2005
2010
2015

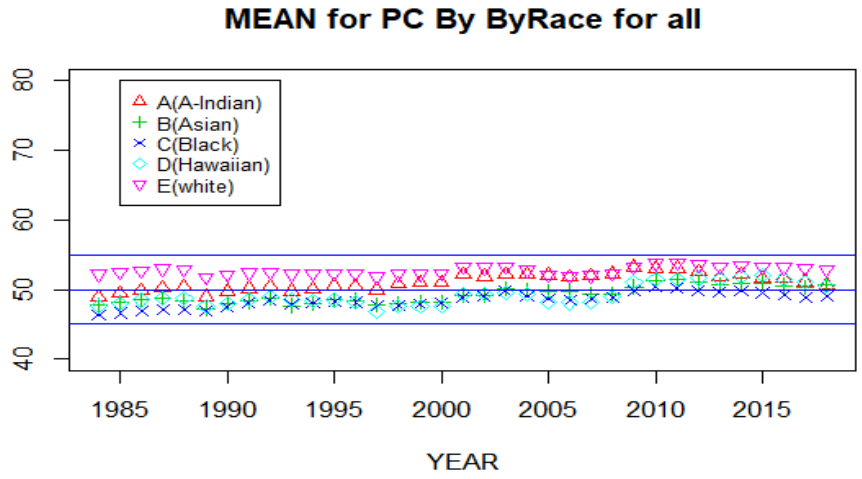

MEAN for MK By ByEthnic for all
A(Hispanic)
B(Non-Hispanic)
D(N/A)
MEAN
YEAR
40
50
60
70
80
1985
1990
1995
2000
2005
2010
2015

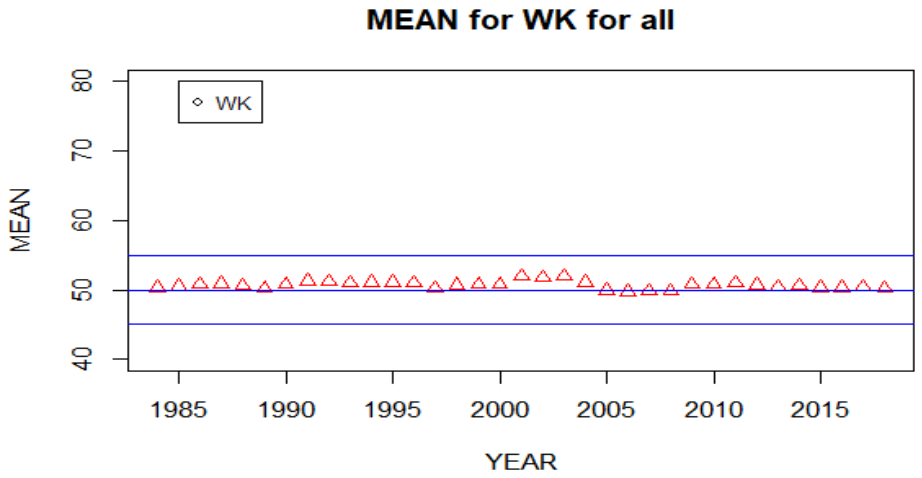

MEAN for MK By ByRace for all
A(A-Indian)
B(Asian)
C(Black)
D(Hawaiian)
E(white)
MEAN
YEAR
40
50
60
70
80
1985
1990
1995
2000
2005
2010
2015

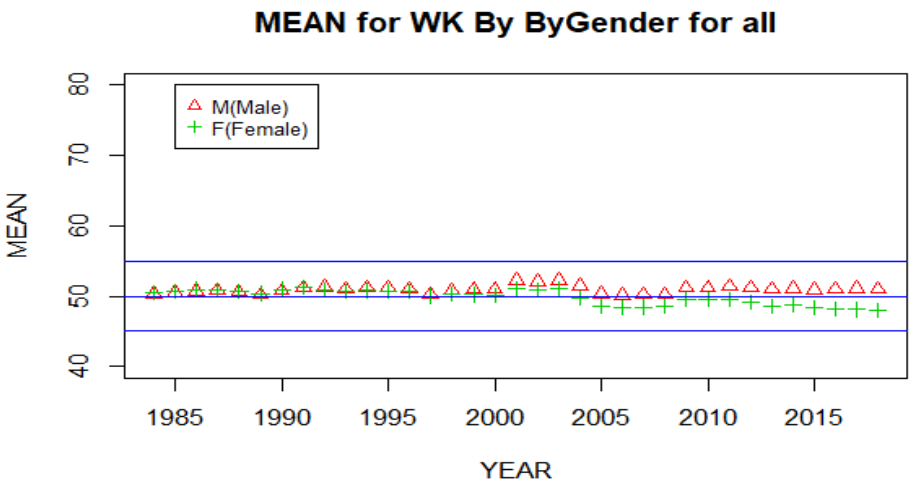

MEAN for PC for all
PC
MEAN
YEAR
40
50
60
70
80
1985
1990
1995
2000
2005
2010
2015

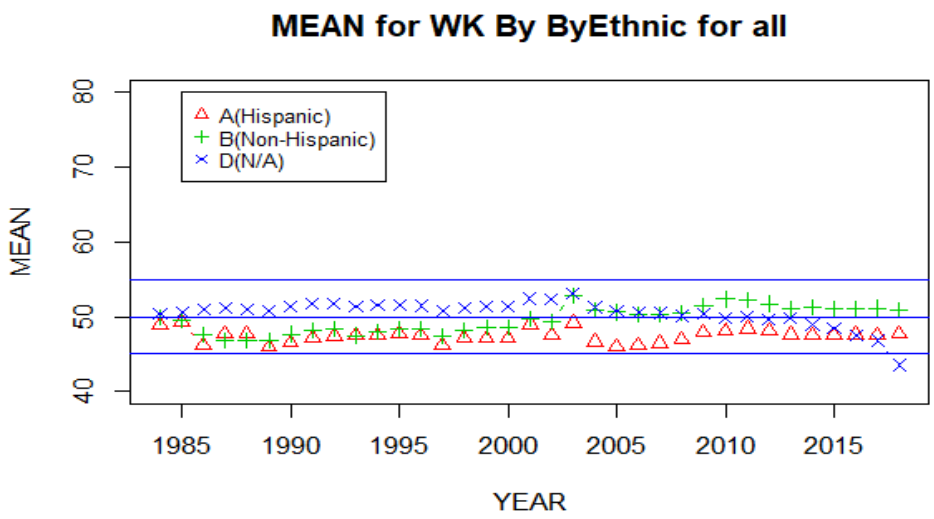

MEAN for PC By ByGender for all
M(Male)
F(Female)
MEAN
YEAR
40
50
60
70
80
1985
1990
1995
2000
2005
2010
2015

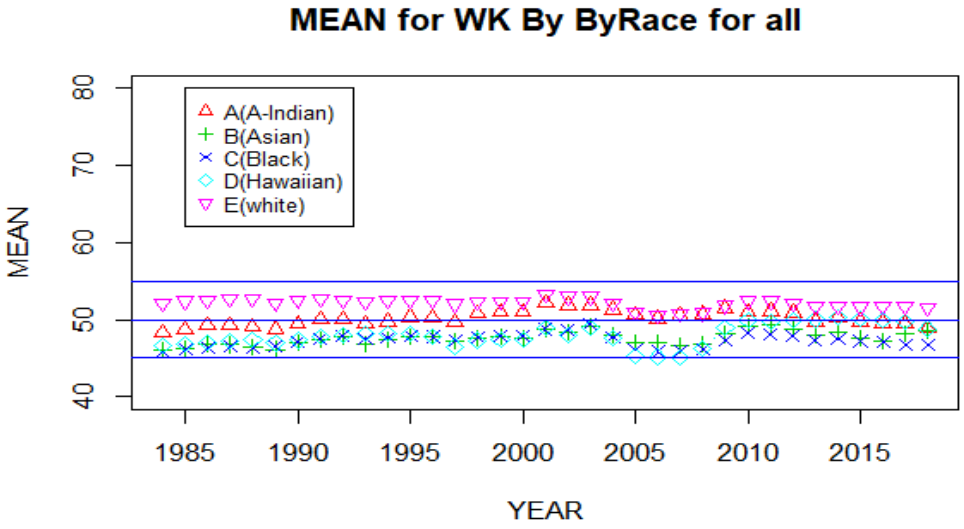

MEAN for PC By ByEthnic for all
A(Hispanic)
B(Non-Hispanic)
D(N/A)
MEAN
YEAR
40
50
60
70
80
1985
1990
1995
2000
2005
2010
2015


MEAN for PC By ByRace for all
A(A-Indian)
B(Asian)
C(Black)
D(Hawaiian)
E(white)
MEAN
YEAR
40
50
60
70
80
1985
1990
1995
2000
2005
2010
2015

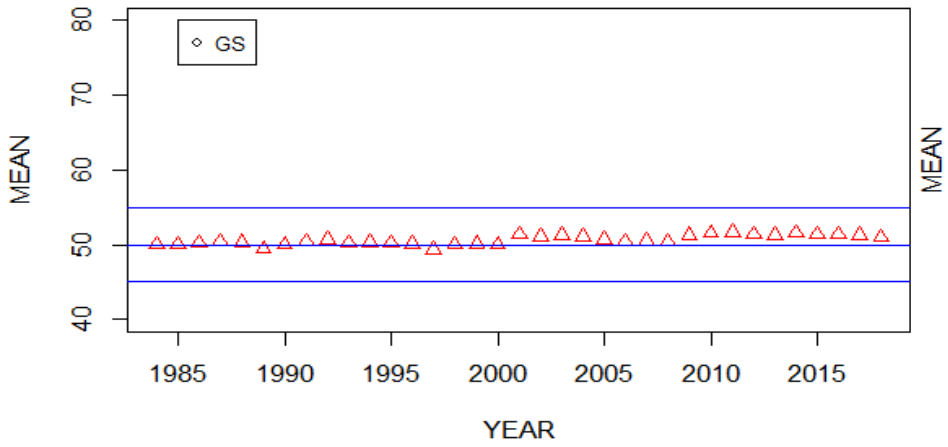

MEAN for WK for all
WK
MEAN
YEAR
40
50
60
70
80
1985
1990
1995
2000
2005
2010
2015


MEAN for WK By ByGender for all
M(Male)
F(Female)
MEAN
YEAR
40
50
60
70
80
1985
1990
1995
2000
2005
2010
2015

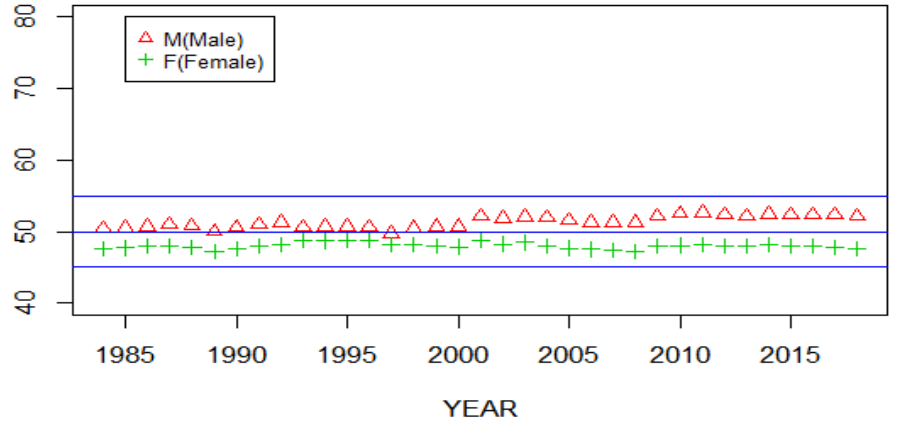

MEAN for WK By ByEthnic for all
A(Hispanic)
B(Non-Hispanic)
D(N/A)
MEAN
YEAR
40
50
60
70
80
1985
1990
1995
2000
2005
2010
2015


MEAN for WK By ByRace for all
A(A-Indian)
B(Asian)
C(Black)
D(Hawaiian)
E(white)
MEAN
YEAR
40
50
60
70
80
1985
1990
1995
2000
2005
2010
2015

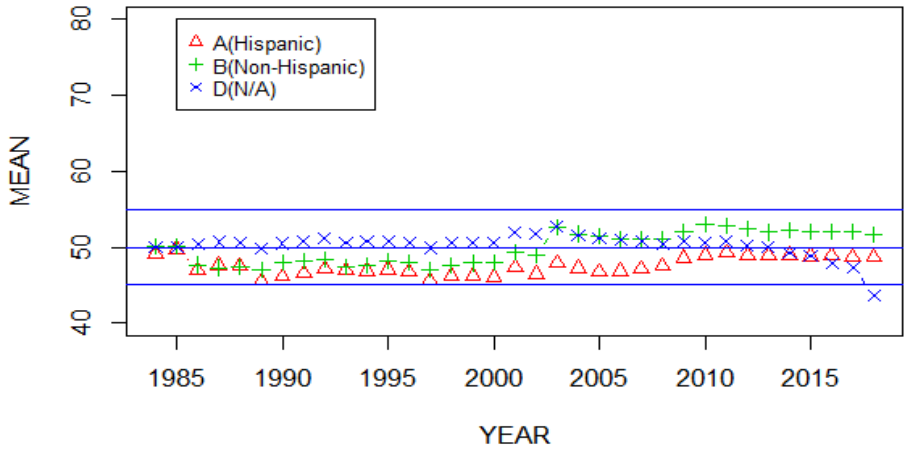
## MEAN for GS for all



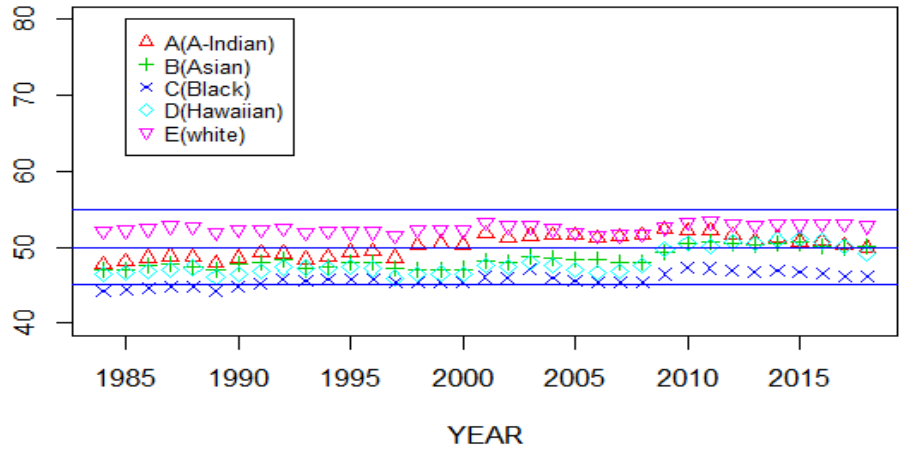

## MEAN for GS By ByGender for all

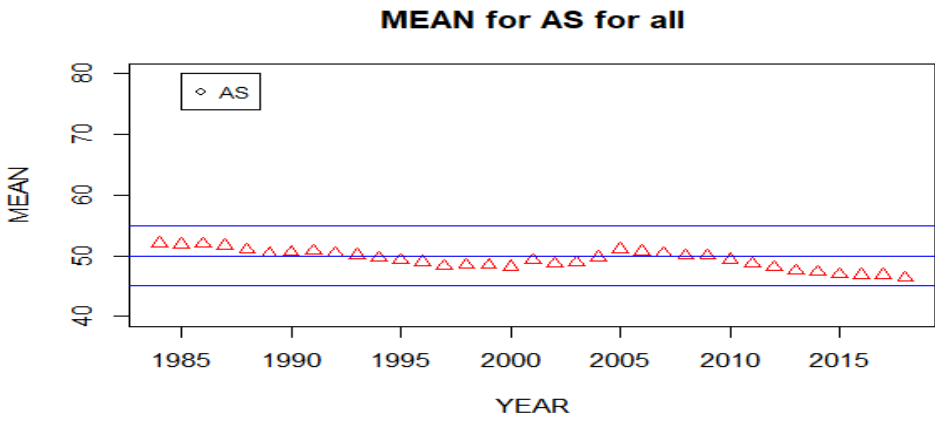


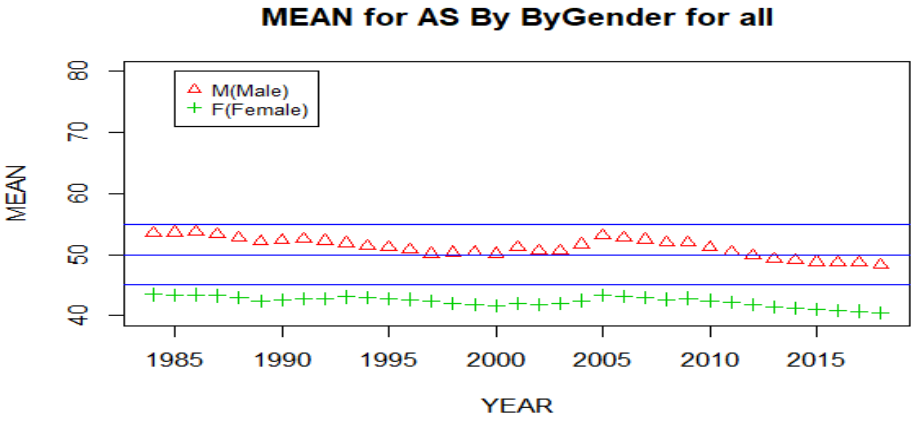

## MEAN for GS By ByEthnic for all

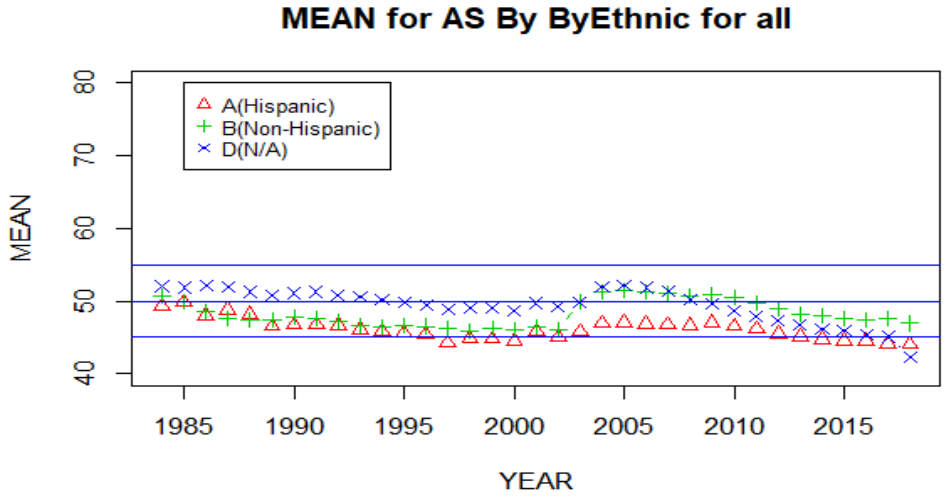


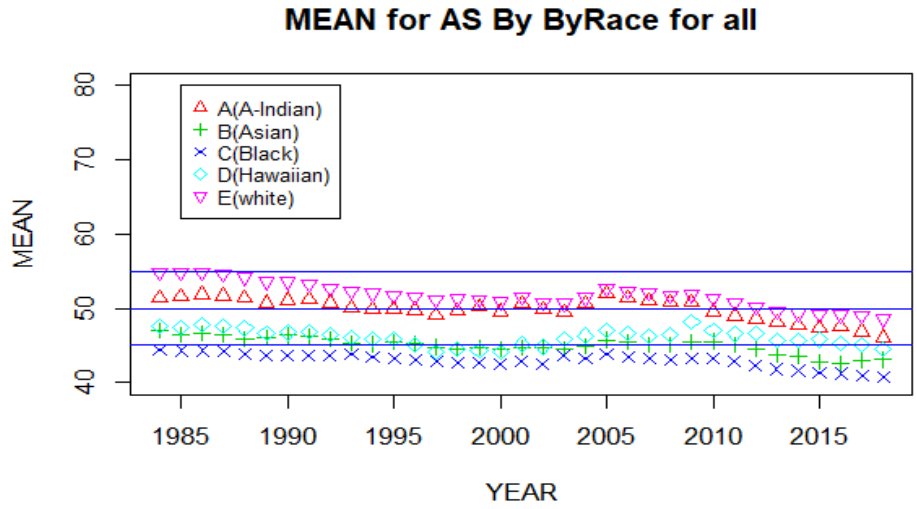

## MEAN for GS By ByRace for all



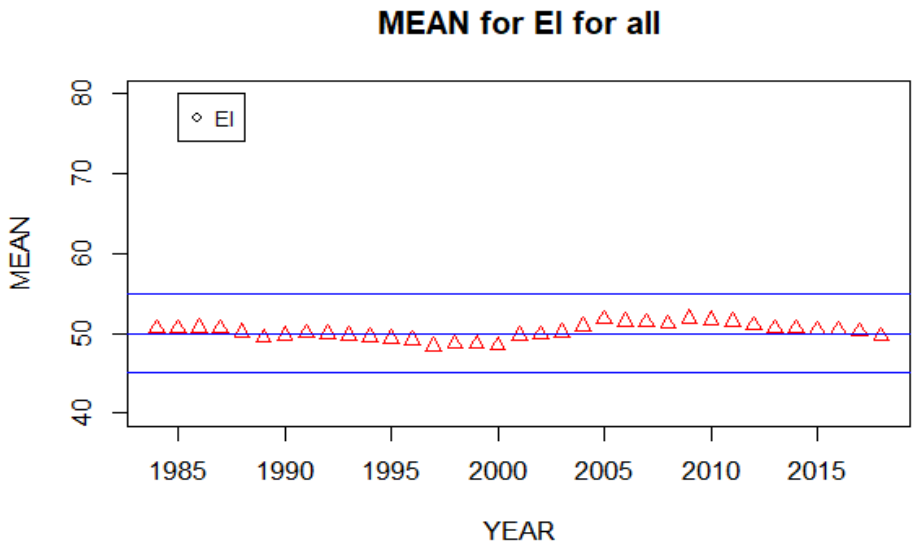


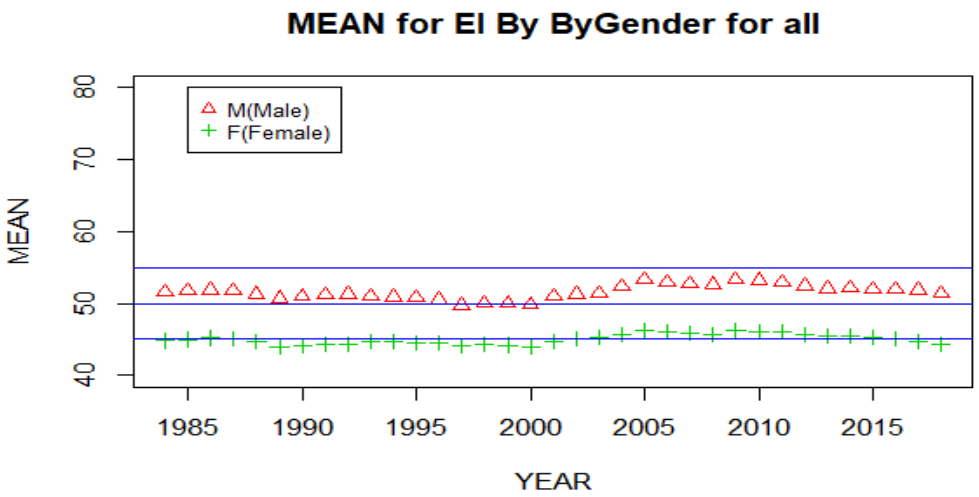



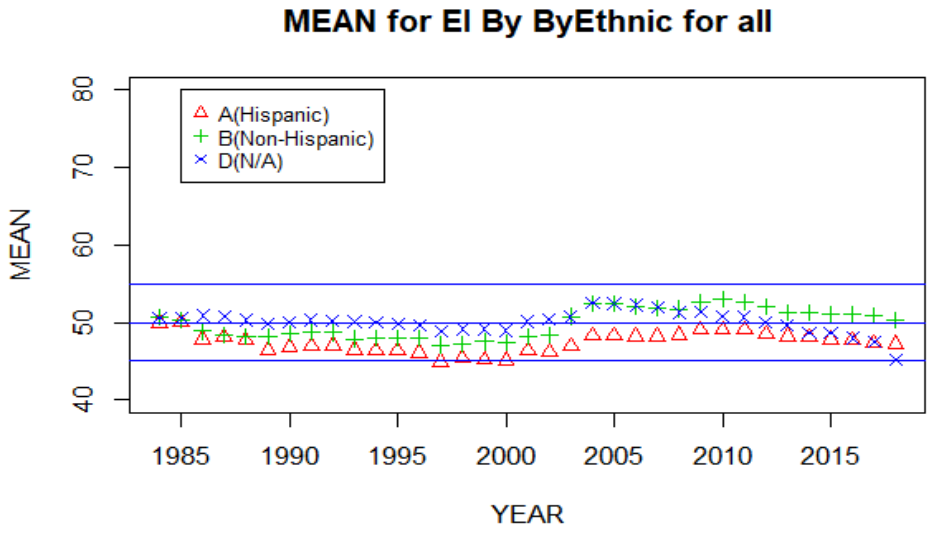



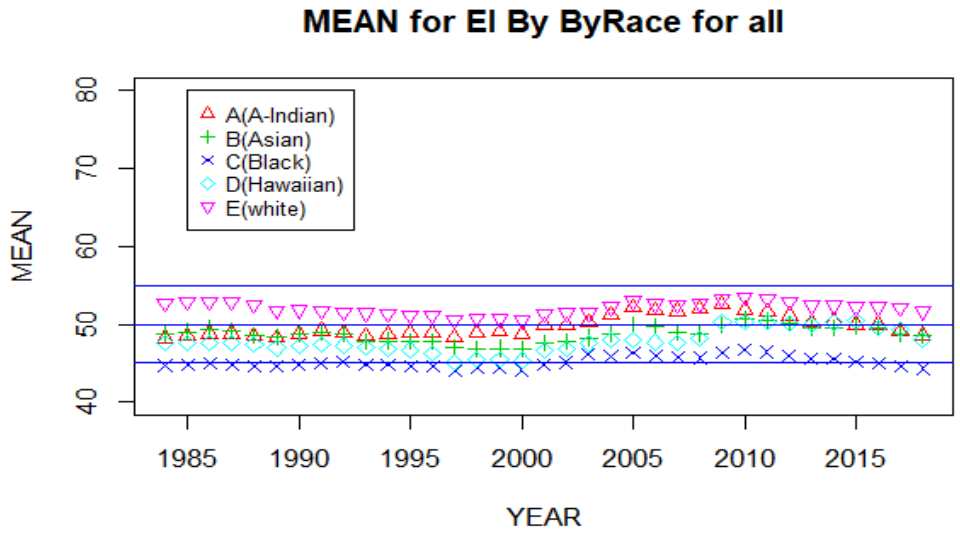



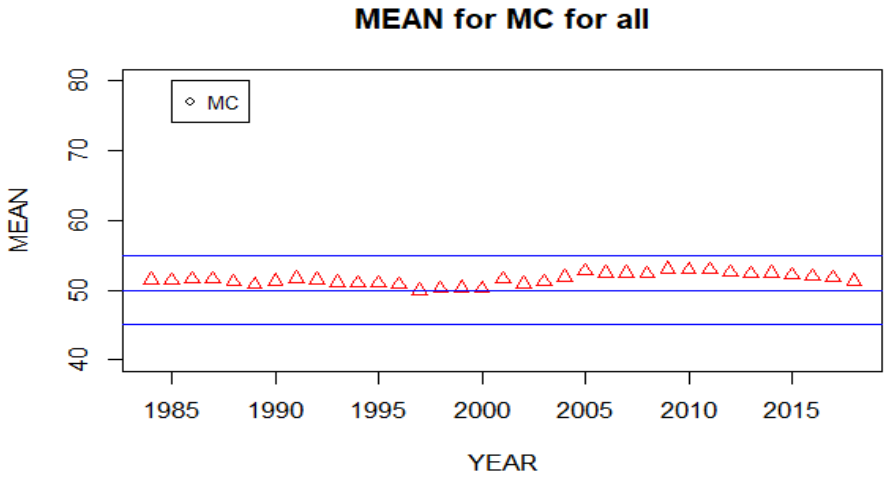

Something notable: AS scores appear to be trending down

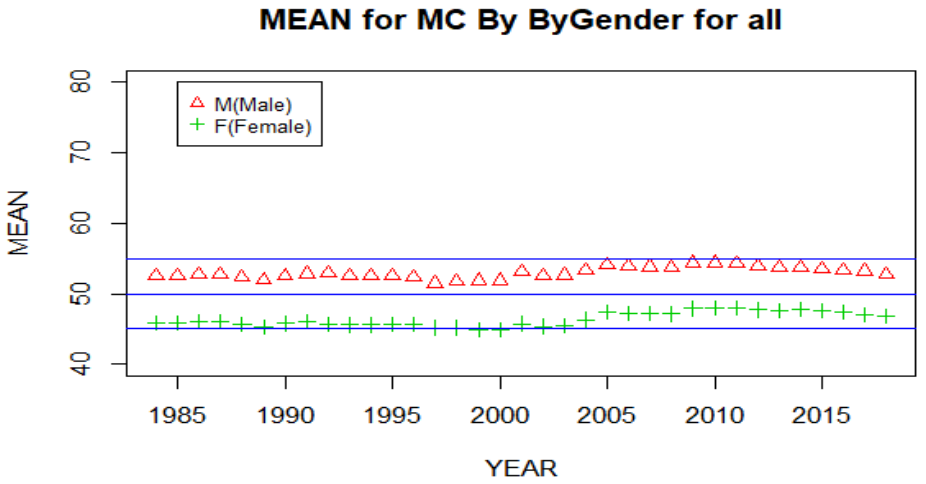

MEAN for EI for all
EI
MEAN
YEAR
40
50
60
70
80
1985
1990
1995
2000
2005
2010
2015

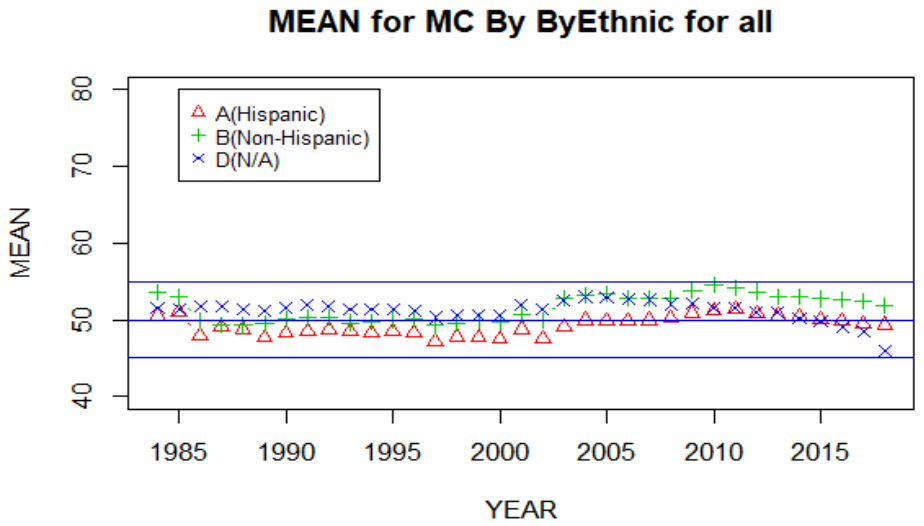

MEAN for EI By ByGender for all
M(Male)
F(Female)
MEAN
YEAR
40
50
60
70
80
1985
1990
1995
2000
2005
2010
2015

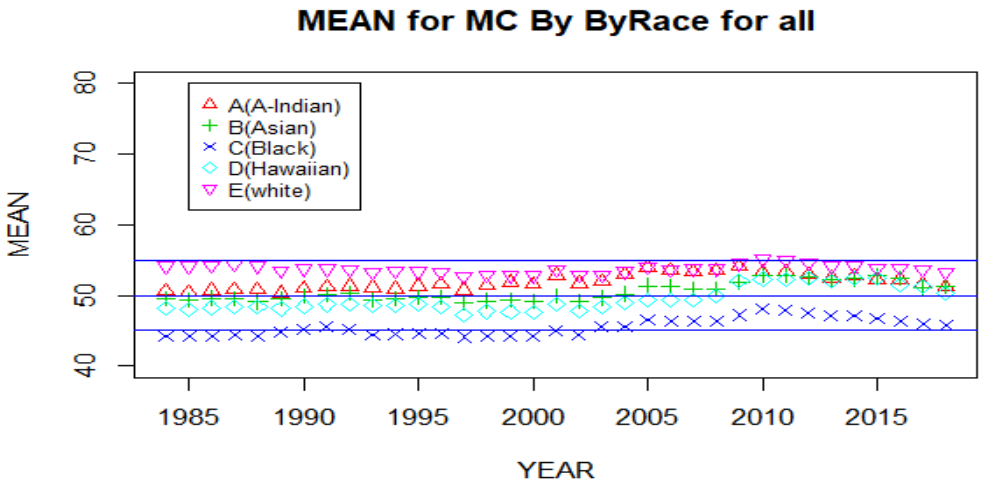

MEAN for EI By ByEthnic for all
A(Hispanic)
B(Non-Hispanic)
D(N/A)
MEAN
YEAR
40
50
60
70
80
1985
1990
1995
2000
2005
2010
2015

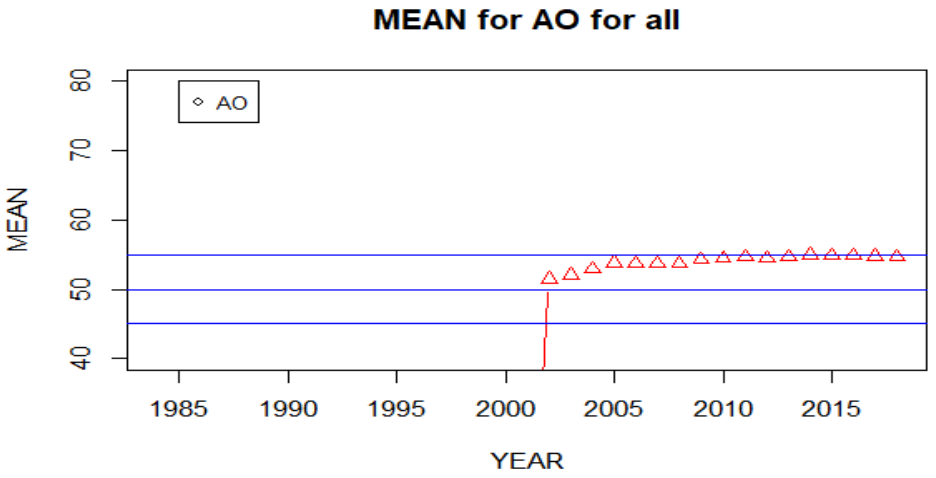

MEAN for EI By ByRace for all
A(A-Indian)
B(Asian)
C(Black)
D(Hawaiian)
E(white)
MEAN
YEAR
40
50
60
70
80
1985
1990
1995
2000
2005
2010
2015

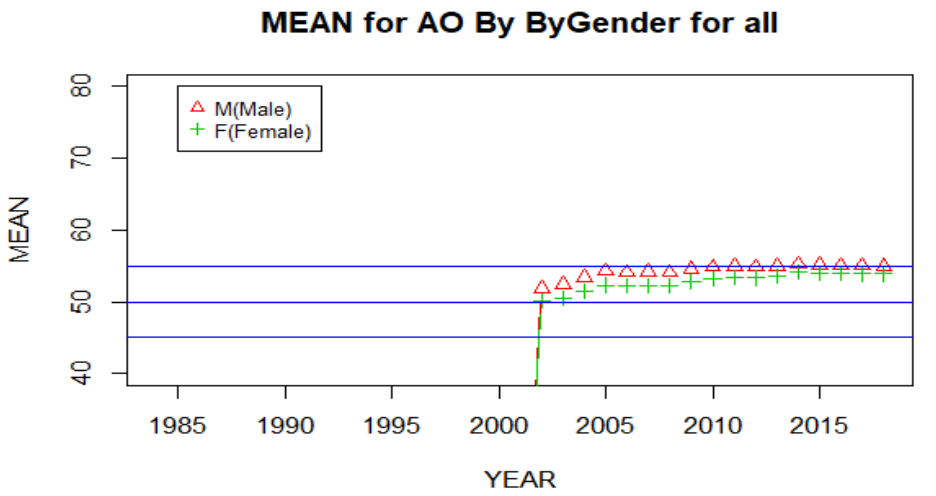

MEAN for MC for all
MC
MEAN
YEAR
40
50
60
70
80
1985
1990
1995
2000
2005
2010
2015

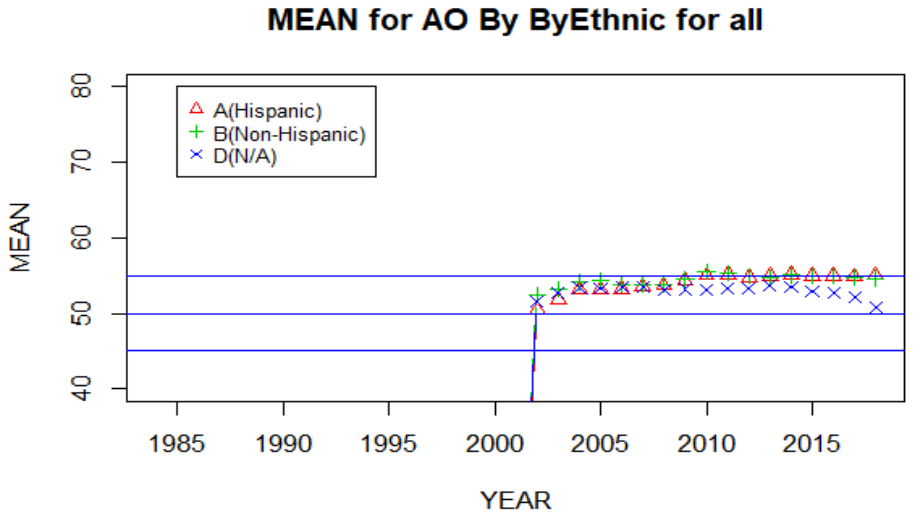

MEAN for MC By ByGender for all
M(Male)
F(Female)
MEAN
YEAR
40
50
60
70
80
1985
1990
1995
2000
2005
2010
2015

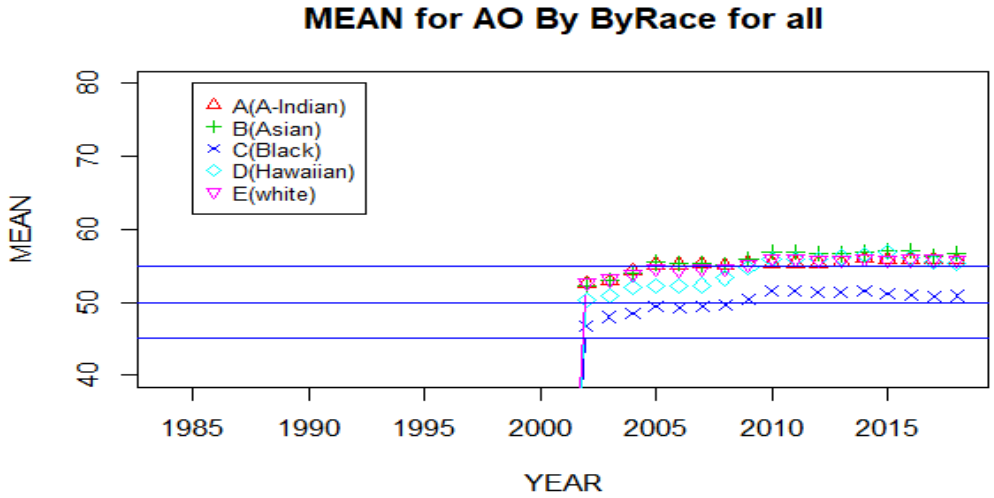

MEAN for MC By ByEthnic for all
A(Hispanic)
B(Non-Hispanic)
D(N/A)
MEAN
YEAR
40
50
60
70
80
1985
1990
1995
2000
2005
2010
2015


MEAN for MC By ByRace for all
A(A-Indian)
B(Asian)
C(Black)
D(Hawaiian)
E(white)
MEAN
YEAR
40
50
60
70
80
1985
1990
1995
2000
2005
2010
2015


MEAN for AO for all
AO
MEAN
YEAR
40
50
60
70
80
1985
1990
1995
2000
2005
2010
2015


MEAN for AO By ByGender for all
M(Male)
F(Female)
MEAN
YEAR
40
50
60
70
80
1985
1990
1995
2000
2005
2010
2015

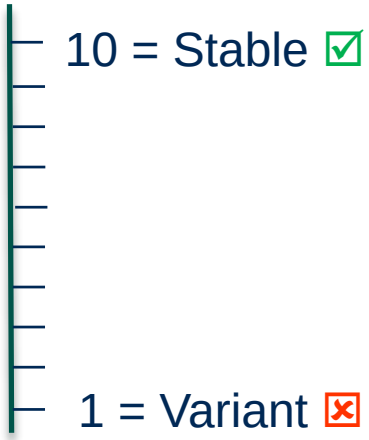

MEAN for AO By ByEthnic for all
A(Hispanic)
B(Non-Hispanic)
D(N/A)
MEAN
YEAR
40
50
60
70
80
1985
1990
1995
2000
2005
2010
2015


MEAN for AO By ByRace for all
A(A-Indian)
B(Asian)
C(Black)
D(Hawaiian)
E(white)
MEAN
YEAR
40
50
60
70
80
1985
1990
1995
2000
2005
2010
2015

# Step 3: Evaluate Usefulness, Appropriateness

- Goal: Evaluate the usefulness and appropriateness of existing tests with regard to the current population.
- Task 3a: Track trends in test scores over years 1984–2018.

| Subtest | Stability Rating |
|---|---|
| General Science | 9 (0.61) |
| Arithmetic Reasoning | 8 (0.89) |
| Word Knowledge | 9 (0.54) |
| Paragraph Comprehension | 8 (0.70) |
| Mathematics Knowledge | 6 (1.44) |
| Electronics Information | 8 (0.93) |
| Automotive Information | 6 (1.63) |
| Shop Information | 6 (1.63) |
| Mechanical Comprehension | 8 (0.81) |
| Assembling Objects | 8 (0.97) |



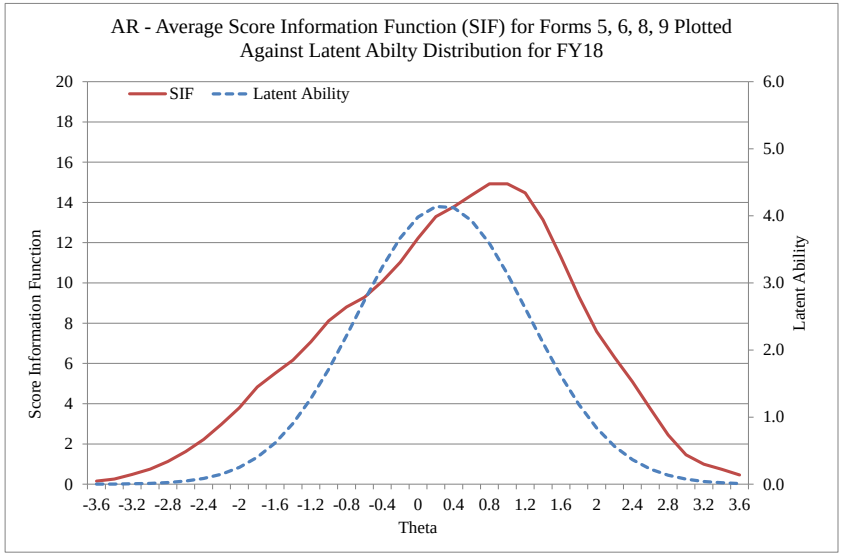
# Step 3: Evaluate Usefulness, Appropriateness

- Goal: Evaluate the usefulness and appropriateness of existing tests with regard to the current population.
- Task 3b: Evaluate what fraction of the population possesses the knowledge/skill assessed by the test.
  - Task 3b(i): Evaluate overlap between latent ability and score information for current testing population.
  - Status: Completed ☑

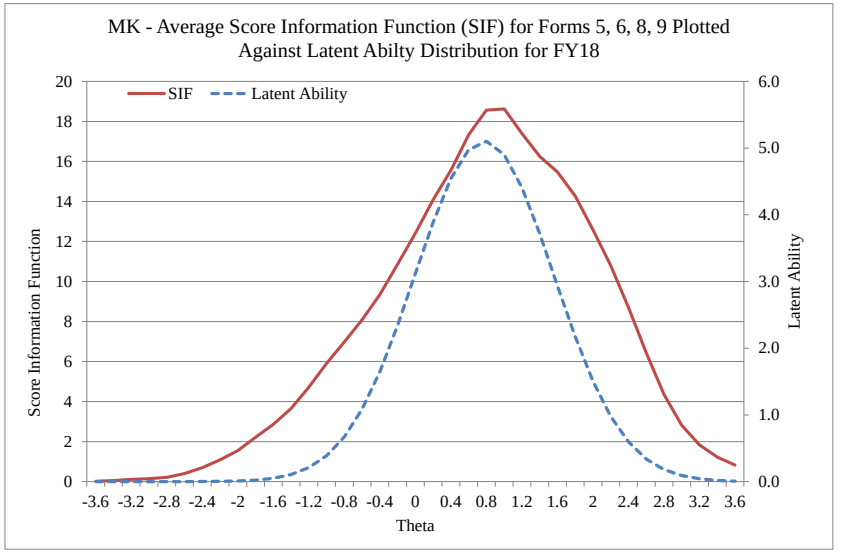
- The latent ability distributions and score information functions appear to be fairly well-aligned for the AFQT tests (i.e., the maximums for the distributions occur at fairly similar abilities for most of the tests).
- The latent ability distributions and score information functions appear to be not so well-aligned for the non-AFQT tests (particularly for AI, SI, EI, and MC).
- The item pools appear to be somewhat more difficult than is needed for the applicant population for all tests except AO, which appears less difficult.

# STEP 3: EVALUATE USEFULNESS, APPROPRIATENESS

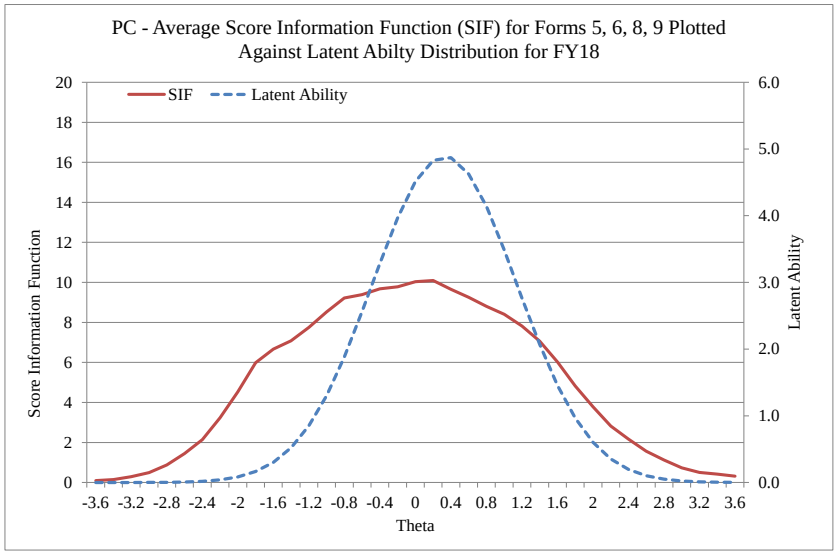
AR



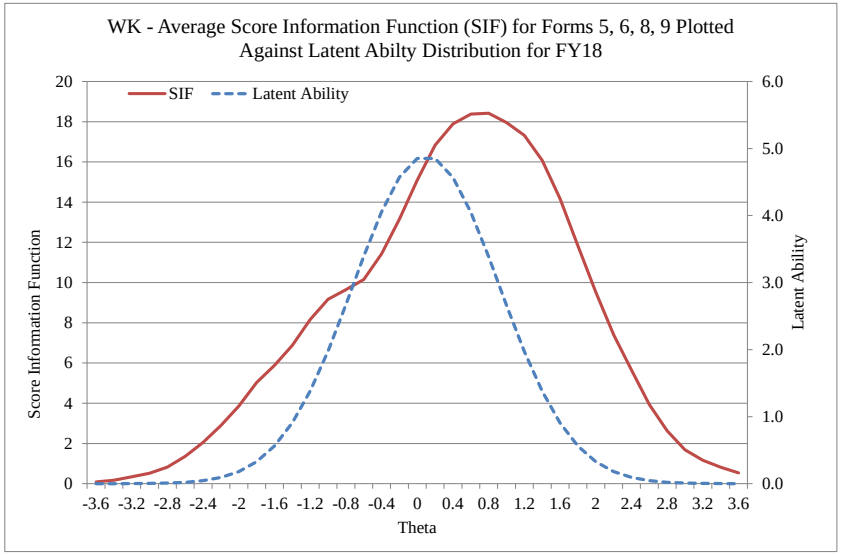

MK




PC

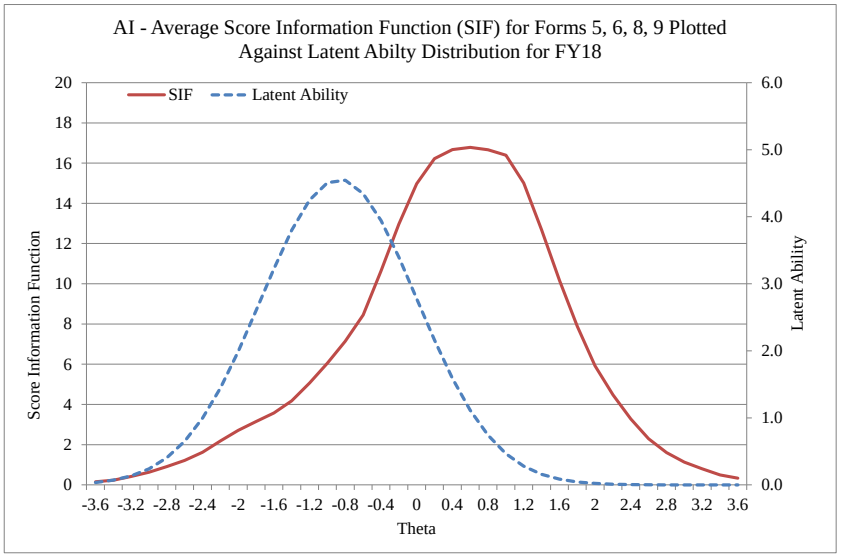



WK

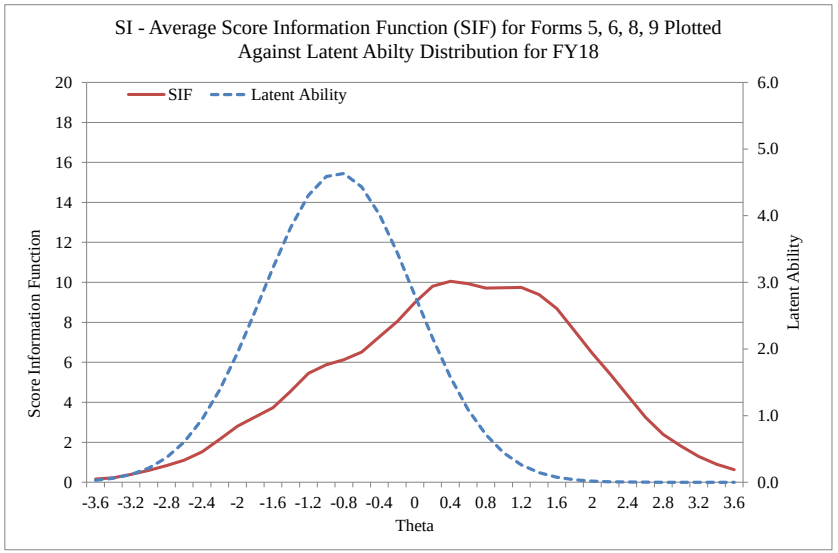


# Step 3: Evaluate Usefulness, Appropriateness

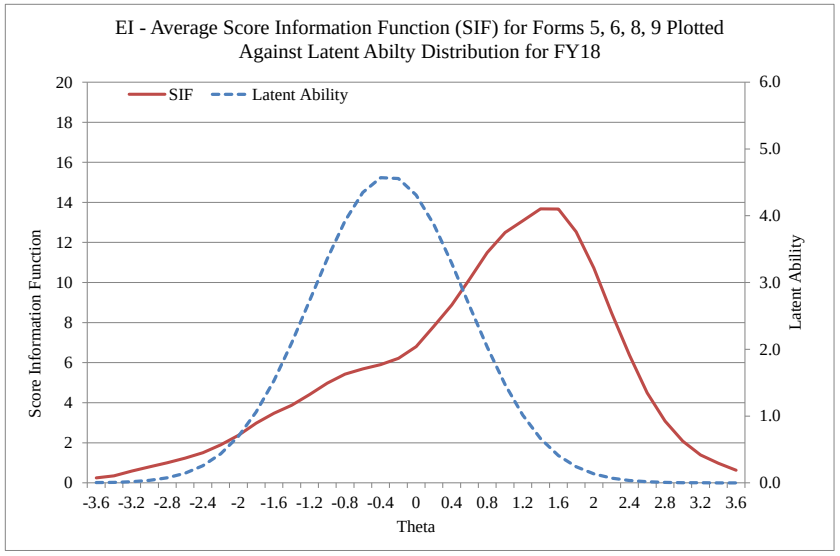
AI



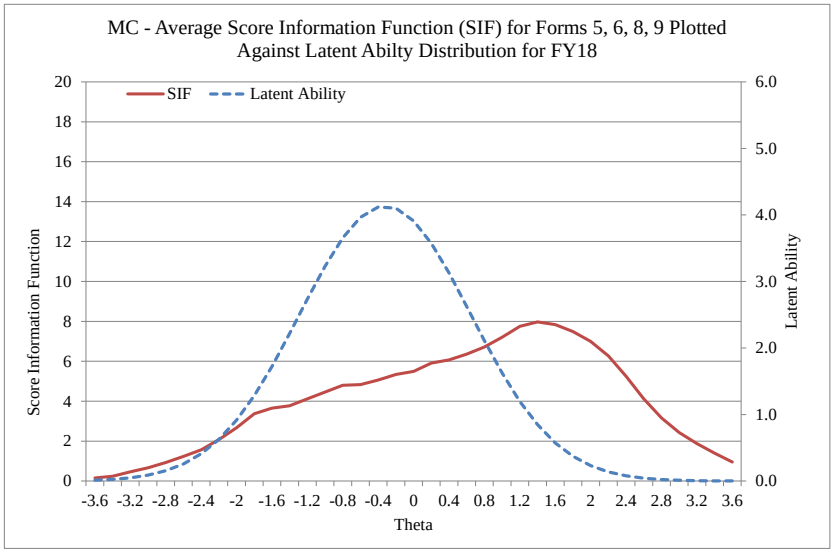

SI




EI

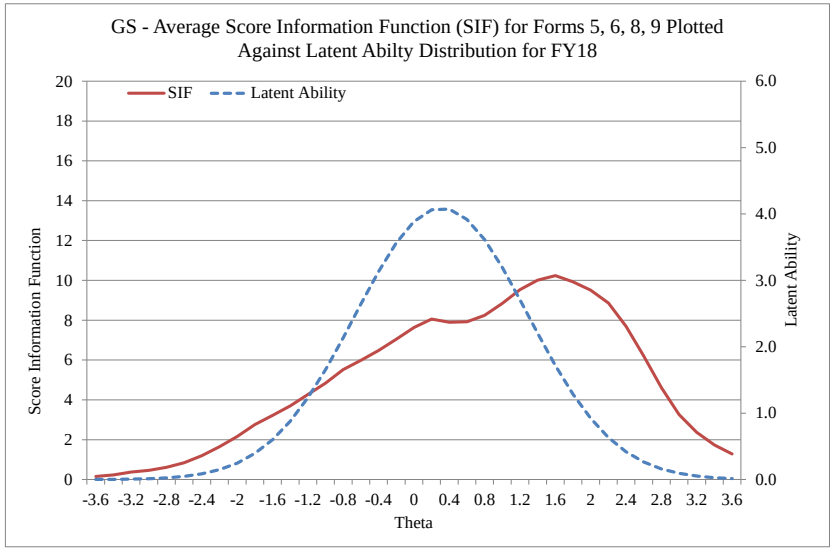


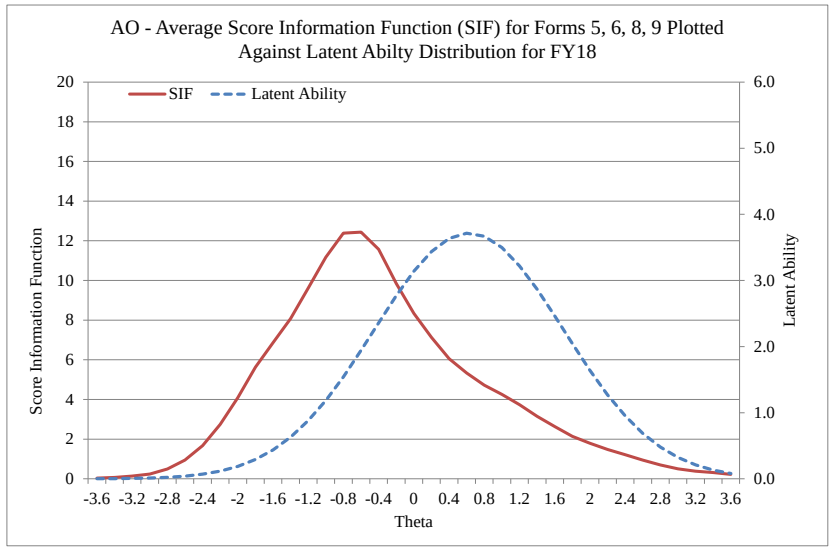

MC



# STEP 3: EVALUATE USEFULNESS, APPROPRIATENESS

GS



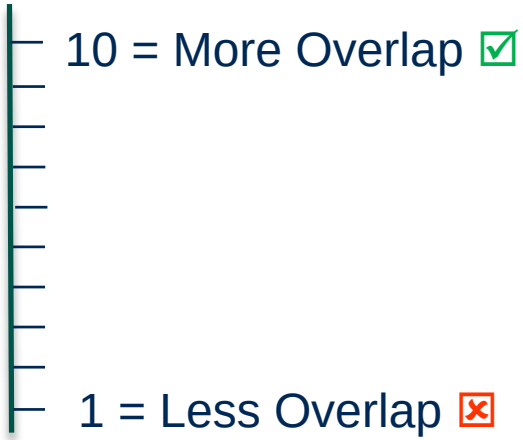




AO

# Step 3: Evaluate Usefulness, Appropriateness

- Goal: Evaluate the usefulness and appropriateness of existing tests with regard to the current population.
- Task 3b: Evaluate what fraction of the population possesses the knowledge/skill assessed by the test.
  - Task 3b(i): Evaluate overlap between latent ability and score information for current testing population.

| Subtest | Finiteness Rating |
|---|---|
| General Science | 7 |
| Arithmetic Reasoning | 9 |
| Word Knowledge | 9 |
| Paragraph Comprehension | 8 |
| Mathematics Knowledge | 9 |
| Electronics Information | 6 |
| Automotive Information | 6 |
| Shop Information | 4 |
| Mechanical Comprehension | 4 |
| Assembling Objects | 6 |



# Step 3: Evaluate Usefulness, Appropriateness

- Goal: Evaluate the usefulness and appropriateness of existing tests with regard to the current population.
  - Task 3b(ii): Use job task analysis ratings to evaluate the relevance of content contained in the science and technical tests to success in technical training.
- Status:

In Progress:
- Development of plan to collect and analyze SME judgments regarding relevance of content in science and technical tests—GS, AI, SI, EI, MC, and Cyber—as prerequisites for success in technical training
- Based on previous ASVAB S&T Training Analysis study (Oppler et al.,1997)

Next Steps:
- Identify jobs to include in data collection
- Identify SMEs to provide job task analysis ratings
- Evaluate relevance of test content to success in technical training

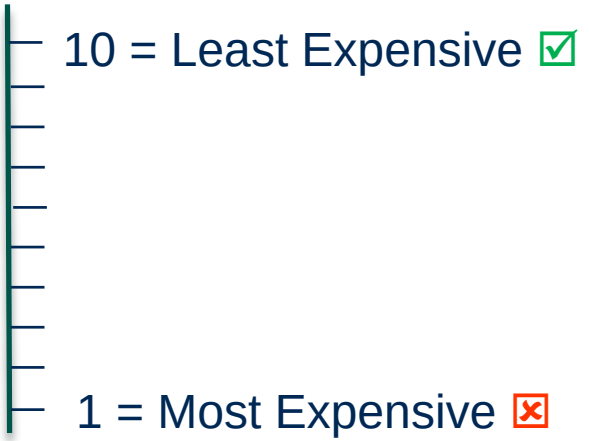
# Step 4: Evaluate Item Development Costs

- Goal: Identify estimated yearly costs for item development.
  - Task 4a: Identify cost per item per test.
  - Task 4b: Identify desired form replacement schedule.
  - Task 4c: Identify number of items needed per year per test.
  - Task 4d: Identify total yearly cost per test.
  - Team: Jeff Harber, Mary Pommerich
  - Status: Completed ☑

# Step 4: Evaluate Item Development Costs

- Goal: Identify estimated yearly costs for item development.

| Subtest | Total Yearly Cost Rating |
|---|---|
| General Science | 3.1 |
| Arithmetic Reasoning | 6.8 |
| Word Knowledge | 2.3 |
| Paragraph Comprehension | 1.0 |
| Mathematics Knowledge | 6.6 |
| Electronics Information | 7.5 |
| Automotive Information | 7.5 |
| Shop Information | 7.5 |
| Mechanical Comprehension | 7.1 |
| Assembling Objects* | TBD |




- The total yearly cost rating is determined as follows:
  - Total yearly cost is computed as the approximate cost per item × target number of pools/items per year and then converted into a rating between 1–10 by dividing by a constant of 27,500 and subtracting from 11 (reverse scoring).

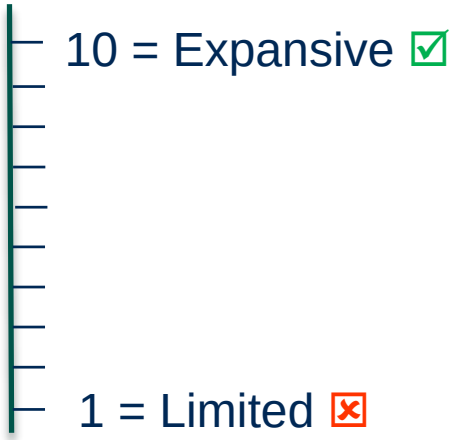
# Step 5: Evaluate Ease of Developing Good Items

- Goal: Evaluate the overall ease/difficulty of developing good quality items.
  - Task 5a: Identify finiteness of domains [limited domain = more difficulty in developing good quality items].
  - Task 5b: Evaluate feasibility of using automatic item generation (AIG) with test content [less feasible = more difficulty in developing good quality items].
  - Task 5c: Identify item retention rates [less retention = more difficulty in developing good quality items].
  - Team: Tia Fechter, Jeff Harber, Mary Pommerich, Matt Trippe
  - Status: Completed ☑

# Step 5: Evaluate Ease of Developing Good Items

- Goal: Evaluate the overall ease/difficulty of developing good quality items.
  - Task 5a: Identify finiteness of domains.

| Subtest | Finiteness Rating |
|---|---|
| General Science | 6 |
| Arithmetic Reasoning | 10 |
| Word Knowledge | 6 |
| Paragraph Comprehension | 10 |
| Mathematics Knowledge | 8 |
| Electronics Information | 4 |
| Automotive Information | 4 |
| Shop Information | 4 |
| Mechanical Comprehension | 5 |
| Assembling Objects | 10 |




The finiteness rating takes into account the following questions:

- How available is content from which to construct new test questions to refresh pool?
- How limited/expansive is the general knowledge for the domain?

# Step 5: Evaluate Ease of Developing Good Items

- Goal: Evaluate the overall ease/difficulty of developing good quality items.
  - Task 5b: Evaluate feasibility of using AIG with test content.

| Subtest | Feasibility of Using AIG Rating† |
|---|---|
| General Science | 4 |
| Arithmetic Reasoning | 8 |
| Word Knowledge | 7 |
| Paragraph Comprehension | 2 |
| Mathematics Knowledge | 8 |
| Electronics Information | 3 |
| Automotive Information | 1 |
| Shop Information | 1 |
| Mechanical Comprehension | 1 |
| Assembling Objects | 10†† |

10 = Highly Feasible ☑

1 = Not Feasible ☒

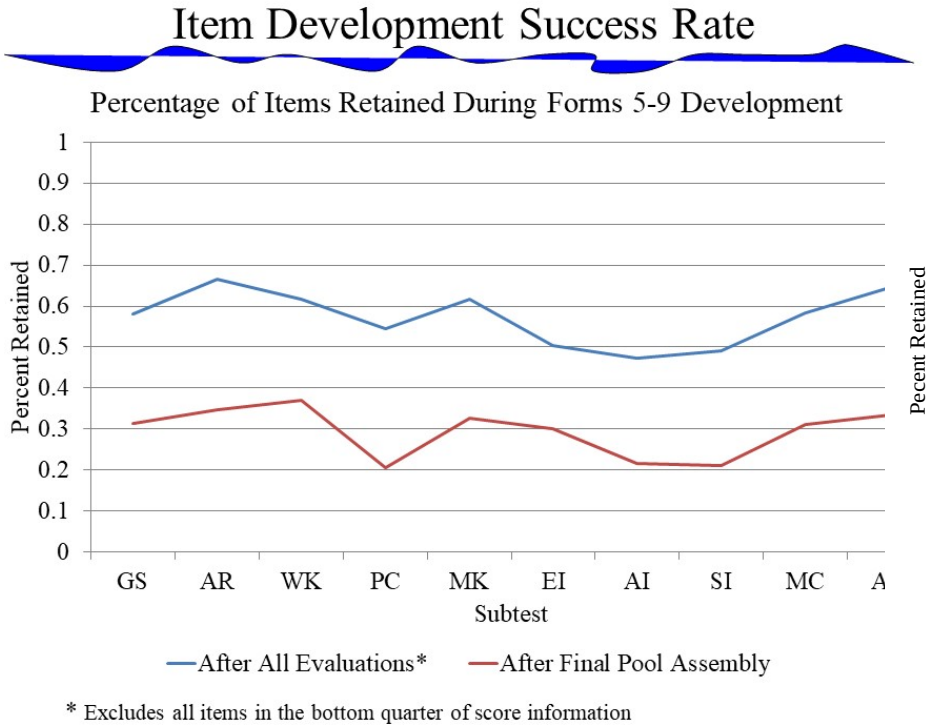
†The feasibility of using AIG rating takes into account the following questions:

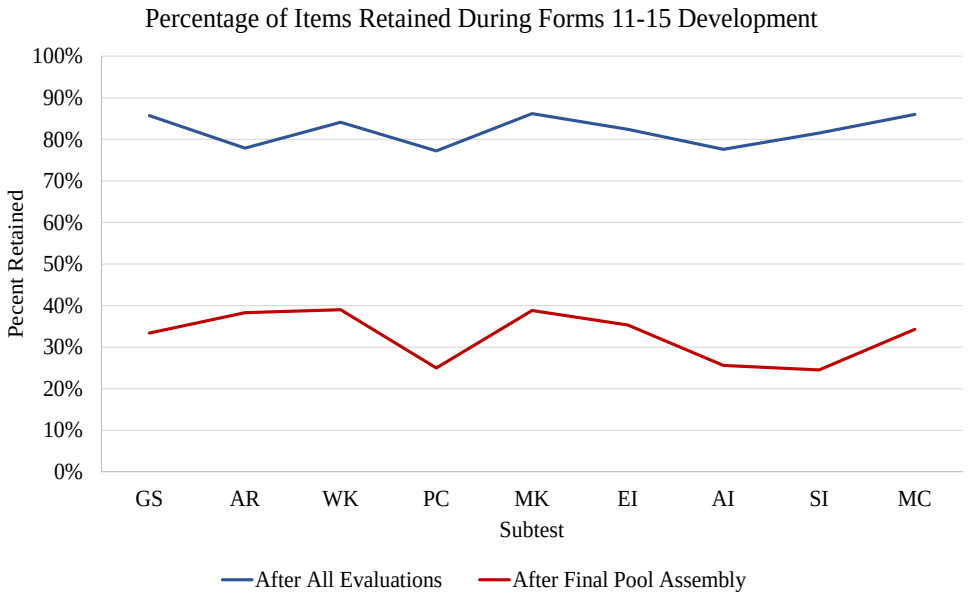
- How much review/formatting/manipulation is required after generation?
- What is the range of content/item types the engine spans for a subtest?
- What is the statistical and content quality of the generated items?
- What percentage of generated items are estimated to be usable?
- Can traditional item tryouts and calibrations be eliminated or requirements reduced?

†† Maximum rating assigned because 5,000+ AO items have already been generated, with no plans to develop additional items in the future. Item tryouts are currently in progress, which will give insight into item quality and the need to revise the rating.

# Step 5: Evaluate Ease of Developing Good Items

- Goal: Evaluate the overall ease/difficulty of developing good quality items.
  - Task 5c: Identify item retention rates.




*Note: The difference in retention rates across Forms 5–9 and Forms 11–15 development for the "After All Evaluations" condition stems from different treatment of low information items (bottom 25% dropped in Forms 5–9 but not in Forms 11–15). This change was triggered by a shift to conducting content and sensitivity reviews during item development, rather than after item evaluations.

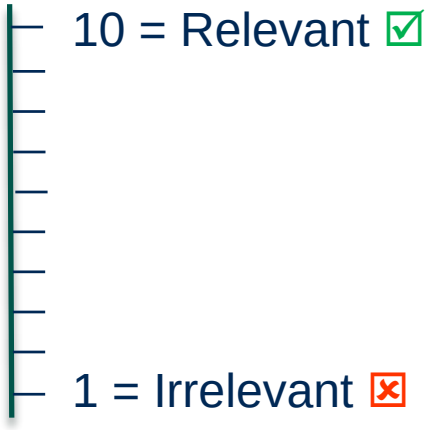
# Step 6: Evaluate Durability of Test Content

- Goal: Evaluate how likely content is to stand the test of time.
- Task 6a: Evaluate extent to which content is (or appears) less relevant to today's applicant population (see also Step 3).
- Task 6b: Evaluate extent to which content is likely to require changes or updates in the near or long term.
  - Consider extent to which content is prone to obsolescence.
  - Consider extent to which content is in need of frequent updating in order to stay current.
  - Consider extent to which it is difficult to keep up with new technology or changes in technology.
- Team: Tia Fechter, Jeff Harber, Sachi Phillips
- Status: Completed ☑

# Step 6: Evaluate Durability of Test Content

- Goal: Evaluate how likely content is to stand the test of time.
- Task 6a: Evaluate extent to which content is (or appears) less relevant to today's applicant population (see also Step 3).

| Subtest | Relevancy Rating |
|---|---|
| General Science | 10 |
| Arithmetic Reasoning | 10 |
| Word Knowledge | 10 |
| Paragraph Comprehension | 10 |
| Mathematics Knowledge | 10 |
| Electronics Information | 6 |
| Automotive Information | 5 |
| Shop Information | 5 |
| Mechanical Comprehension | 8 |
| Assembling Objects | N/A |



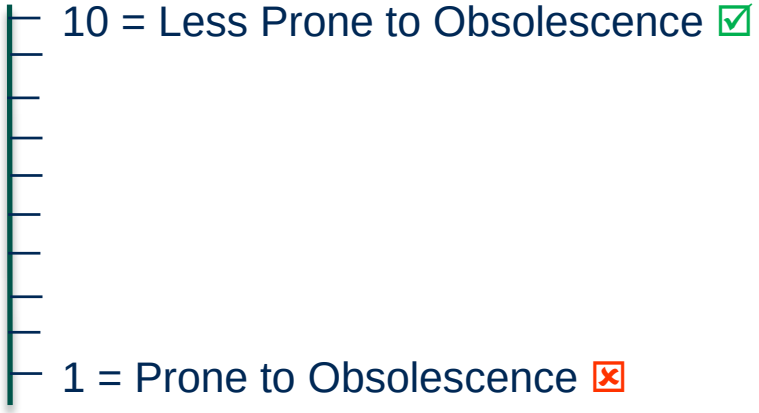

The relevancy rating takes into account the following question:

- Is the domain's content currently emphasized within high schools?

# Step 6: Evaluate Durability of Test Content

- Goal: Evaluate how likely content is to stand the test of time.
- Task 6b: Evaluate extent to which content is likely to require changes or updates in the near or long term.

| Subtest | Obsolescence Rating |
|---|---|
| General Science | 7 |
| Arithmetic Reasoning | 10 |
| Word Knowledge | 10 |
| Paragraph Comprehension | 10 |
| Mathematics Knowledge | 10 |
| Electronics Information | 6 |
| Automotive Information | 3 |
| Shop Information | 8 |
| Mechanical Comprehension | 10 |
| Assembling Objects | 10 |




The obsolescence rating takes into account the following questions:

- To what extent is the domain's content vulnerable to obsolescence?
- To what extent is the domain's content in need of frequent updating to stay current?
- To what extent is new technology or changes in technology impacting the domain?

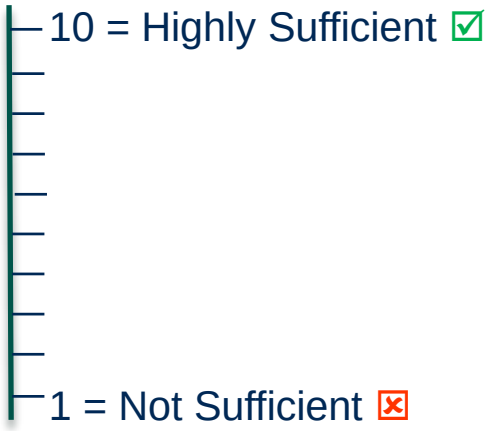
# Step 7: Evaluate Efficiency of Content Coverage

- Goal: Review prior research and summarize findings regarding the efficiency and adequacy of content coverage (i.e., redundancies and gaps).
  - Consider redundancies in content coverage across tests.
  - Consider gaps in content coverage.
  - Consider potentially unnecessary content coverage.
- Team: Tia Fechter, Jeff Harber
- Status: Completed ☑

# Step 7: Evaluate Efficiency of Content Coverage

- Goal: Review prior research and summarize findings regarding the efficiency and adequacy of content coverage (i.e., redundancies and gaps).

| Subtest | Content Efficiency Rating | Recommendation Summary |
|---|---|---|
| General Science** | 7 | Add content areas |
| Arithmetic Reasoning* | 8 | Drop content areas |
| Word Knowledge | 10 | |
| Paragraph Comprehension | 9 | |
| Mathematics Knowledge* | 7 | Drop content areas |
| Electronics Information** | 7 | Add content areas |
| Automotive Information | 9 | |
| Shop Information | 9 | |
| Mechanical Comprehension** | 7 | |
| Assembling Objects | 10 | |




*Prior research suggests merging subtests could be possible (also being investigated in Step 14)

**The feasibility of merging subtests being investigated (see Step 16.5)

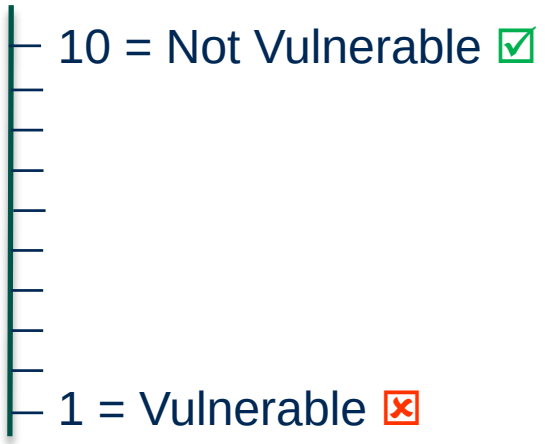
# Step 8: Evaluate Vulnerability to Compromise

- Goal: Evaluate the vulnerability of item content and item pools to compromise.
  - Consider features of tests that could make them easy to compromise.
  - Consider features of item pools that could make them easy to compromise.
  - Consider previous incidences of compromise on the ASVAB and tests that were breached.
- Team: Tia Fechter, Jeff Harber, Sachi Phillips, Dan Segall
- Status: Completed ☑

# Step 8: Evaluate Vulnerability to Compromise

- Goal: Evaluate the vulnerability of item content and item pools to compromise.

| Subtest | Vulnerability Rating |
|---|---|
| General Science | 7 |
| Arithmetic Reasoning | 6 |
| Word Knowledge | 2 |
| Paragraph Comprehension | 7 |
| Mathematics Knowledge | 6 |
| Electronics Information | 8 |
| Automotive Information | 7 |
| Shop Information | 7 |
| Mechanical Comprehension | 8 |
| Assembling Objects | 10 |




The vulnerability rating takes into account the following questions:

- What are the stakes for performing well (e.g., determine selection, incentive eligibility)?
- What are the benefits of piracy at an individual and organization level?
- Are there features of the tests/pools that make them more susceptible to compromise?

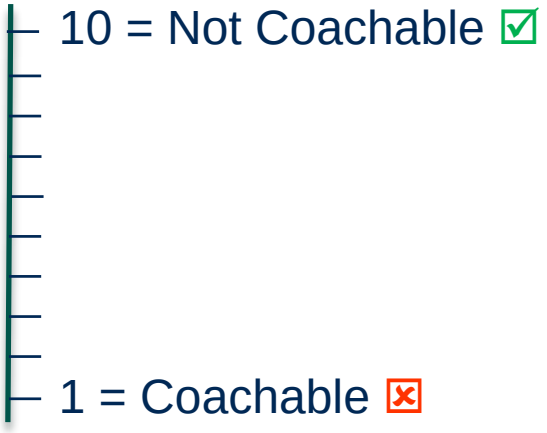
# Step 9: Evaluate Other Vulnerabilities

- Goal: Evaluate the vulnerability of item content to other unwanted effects.
- Task 9a: Coachability
- Task 9b: Practice Effects
- Task 9c: Hardware Effects
- Task 9d: Mode Effects
- Task 9e: Local Dependence
- Task 9f: Device Familiarity
- Team: Tia Fechter, Jeff Harber, Sachi Phillips, Mary Pommerich, Dan Segall
- Status: Tasks 9a–9e completed ☑

# Step 9: Evaluate Other Vulnerabilities

- Goal: Evaluate the vulnerability of item content to other unwanted effects.
- Task 9a: Coachability

| Subtest | Coachability Rating |
|---|---|
| General Science | 10 |
| Arithmetic Reasoning | 10 |
| Word Knowledge | 9 |
| Paragraph Comprehension | 8 |
| Mathematics Knowledge | 10 |
| Electronics Information | 10 |
| Automotive Information | 10 |
| Shop Information | 10 |
| Mechanical Comprehension | 8 |
| Assembling Objects | 5 |




The coachability rating takes into account the following question:

- How susceptible is the domain's content or test's format to score increases due to learned test-taking techniques rather than increased knowledge of course material?

# Step 9: Evaluate Other Vulnerabilities

- Goal: Evaluate the vulnerability of item content to other unwanted effects.
- Task 9b: Practice Effects

| Subtest | Practice Effect Rating |
|---|---|
| General Science | 10 |
| Arithmetic Reasoning | 6 |
| Word Knowledge | 10 |
| Paragraph Comprehension | 10 |
| Mathematics Knowledge | 6 |
| Electronics Information | 7 |
| Automotive Information | 10 |
| Shop Information | 10 |
| Mechanical Comprehension | 7 |
| Assembling Objects | 1 |

10 = Not Easy to Increase Score with Practice ☑

1 = Easy to Increase Score with Practice ☒

The practice effect rating takes into account the following question:

- Could taking the test multiple times result in improved scores, without learning between test occasions?

# Step 9: Evaluate Other Vulnerabilities

- Goal: Evaluate the vulnerability of item content to other unwanted effects.
- Task 9c: Hardware Effects

| Subtest | Hardware Effect Rating* |
|---|---|
| General Science | 10 |
| Arithmetic Reasoning | 6 |
| Word Knowledge | 10 |
| Paragraph Comprehension | 10 |
| Mathematics Knowledge | 6 |
| Electronics Information | 10 |
| Automotive Information | 10 |
| Shop Information | 10 |
| Mechanical Comprehension | 5 |
| Assembling Objects | 4 |

10 = Not Susceptible to Score Differences/ Response Time Differences ☑

1 = Susceptible to Score Differences/Response Time Differences ☒

The hardware effect rating takes into account the following questions:

- Are there differences in hardware across test-taking environments that would enhance or hamper performance?
- Are there differences in hardware (e.g., monitor size) across test-taking environments that would enhance or hamper test completion rates?

*Observed score differences for individual subtests are sometimes only for specific forms. Once forms are equated, there are no score differences that impact qualification rates for the respective composites that the subtests are a part of.

# Step 9: Evaluate Other Vulnerabilities

- Goal: Evaluate the vulnerability of item content to other unwanted effects.
- Task 9d: Mode Effects

| Subtest | Mode Effect Rating* |
|---|---|
| General Science | 10 |
| Arithmetic Reasoning | 10 |
| Word Knowledge | 10 |
| Paragraph Comprehension | 5 |
| Mathematics Knowledge | 8 |
| Electronics Information | 5 |
| Automotive Information | 6 |
| Shop Information | 6 |
| Mechanical Comprehension | 10 |
| Assembling Objects | 9 |

10 = Difference between P&P and CAT Unlikely ☑

1 = Difference between P&P and CAT Likely ☒

The mode effect rating takes into account the following questions:

- Are there differences in test modes (e.g., P&P vs. CAT) that would enhance or hamper performance?
- Are there differences in test modes (e.g., P&P vs. CAT) that would enhance or hamper test completion rates?

*Observed score differences for individual subtests are only for specific subgroups. Once subtest scores are combined within composites (e.g., AFQT), there are no score differences that impact qualification rates.

# Step 9: Evaluate Other Vulnerabilities

- Goal: Evaluate the vulnerability of item content to other unwanted effects.
- Task 9e: Local Dependence

| Subtest | Local Dependence Rating |
|---|---|
| General Science | 7 |
| Arithmetic Reasoning | 9 |
| Word Knowledge | 9 |
| Paragraph Comprehension | 9 |
| Mathematics Knowledge | 5 |
| Electronics Information | 7 |
| Automotive Information | 7 |
| Shop Information | 7 |
| Mechanical Comprehension | 5 |
| Assembling Objects | 10 |

10 = Local Dependencies Unlikely ☑

1 = Local Dependencies Likely ☒

The local dependence rating takes into account the following questions:

- How susceptible is the test to local dependence issues?
- How much effort goes into identifying item enemies?
- Are item selection controls in place to limit the impact of local dependence?

# Step 9: Evaluate Other Vulnerabilities

- Goal: Evaluate the vulnerability of item content to other unwanted effects.
- Task 9f: Device Familiarity

| Subtest | Familiarity Effect Rating |
|---|---|
| General Science | 6 |
| Arithmetic Reasoning | 6 |
| Word Knowledge | 10 |
| Paragraph Comprehension | 8 |
| Mathematics Knowledge | 7 |
| Electronics Information | 8 |
| Automotive Information | 8 |
| Shop Information | 8 |
| Mechanical Comprehension | 8 |
| Assembling Objects | 6 |

10 = Device Familiarity Does Not Significantly Impact Performance/Response Time ☑

1 = Device Familiarity Significantly Impacts Performance/Response Time ☒

# Step 10: Evaluate Efficiency of Each Test

- Goal: Evaluate the relative efficiency of each test with regard to testing time allotted and testing time used.
- Task 10a: Summarize total testing time allocated on CAT-ASVAB.
- Task 10b: Summarize observed testing times for applicants, total and per test.
- Task 10c: Summarize time allocated versus time spent, per item and per test.
- Team: Furong Gao, Mary Pommerich, Dan Segall
- Status: Completed ☑

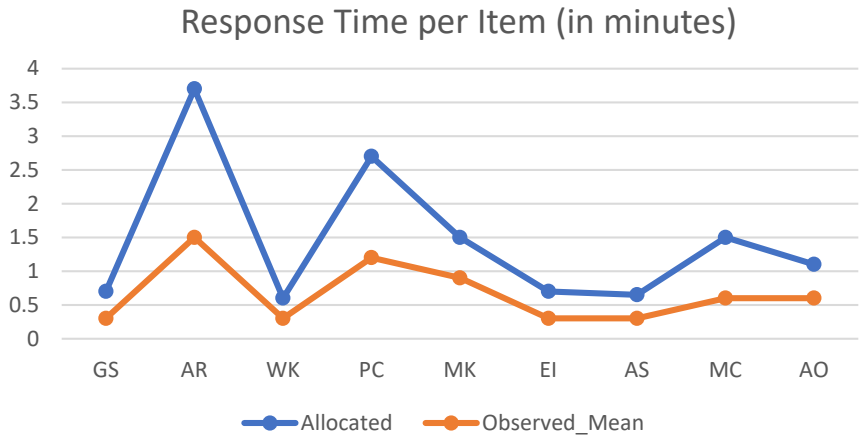
# Step 10: Evaluate Efficiency of Each Test

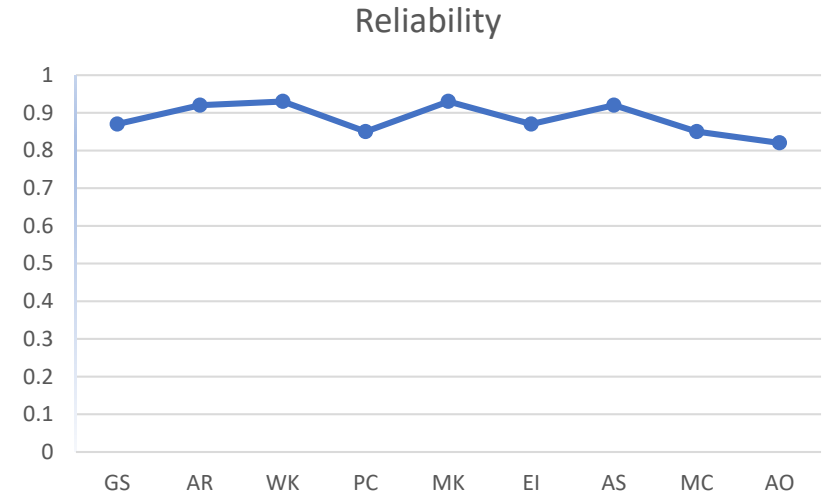
- Goal: Evaluate the relative efficiency of each test with regard to testing time allotted and testing time used.
- Task 10a–10c: Test time summary (per item stats are in parentheses)

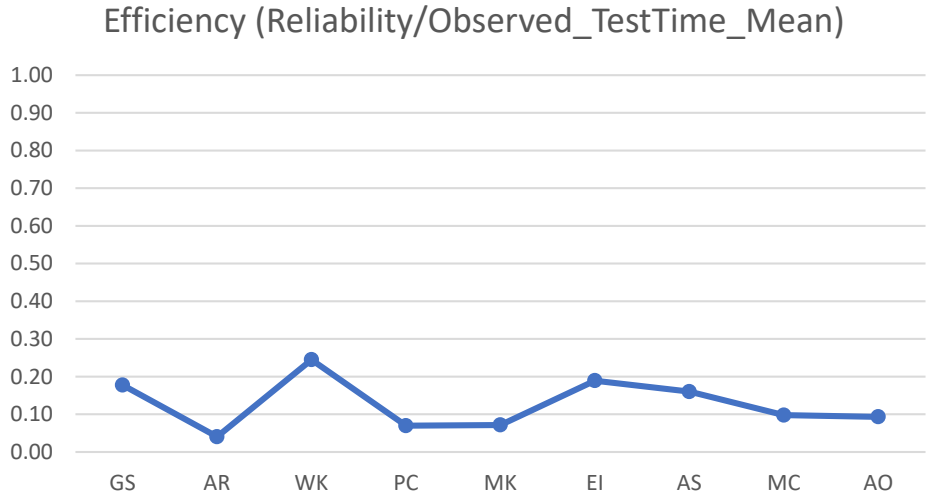
| | | | Allocated | | Observed[2] (mean) | |
|---|---|---|---|---|---|---|
| Subtest | # of Scored Questions | Reliability[1] | Without Seed | With Seed | Without Seed | With Seed |
| GS | 15 | 0.87 | 10 (0.7) | 20 (0.7) | 5 (0.3) | 10 (0.3) |
| AR | 15 | 0.92 | 55 (3.7) | 113 (3.8) | 23 (1.5) | 48 (1.6) |
| WK | 15 | 0.93 | 9 (0.6) | 18 (0.6) | 4 (0.3) | 7 (0.2) |
| PC | 10 | 0.85 | 27 (2.7) | 75 (3.0) | 12 (1.2) | 34 (1.4) |
| MK | 15 | 0.93 | 23 (1.5) | 47 (1.6) | 13 (0.9) | 28 (0.9) |
| EI | 15 | 0.87 | 10 (0.7) | 21 (0.7) | 5 (0.3) | 10 (0.3) |
| AS | 20 | 0.92 | 13 (0.7) | 33/28 (0.9/0.8)[3] | 6 (0.3) | 14/12 (0.7/0.6)[3] |
| MC | 15 | 0.85 | 22 (1.5) | 42 (1.4) | 9 (0.6) | 16 (0.5) |
| AO | 15 | 0.82 | 17 (1.1) | 36 (1.2) | 9 (0.6) | 19 (0.6) |
| **Total** | **135** | | **186 (1.4)** | | **85 (0.6)** | |

[1] From www.officialasvab.com
[2] 2015-2018 WinCAT and iCAT data
[3] With AI/SI seeded items

# Step 10: Evaluate Efficiency of Each Test



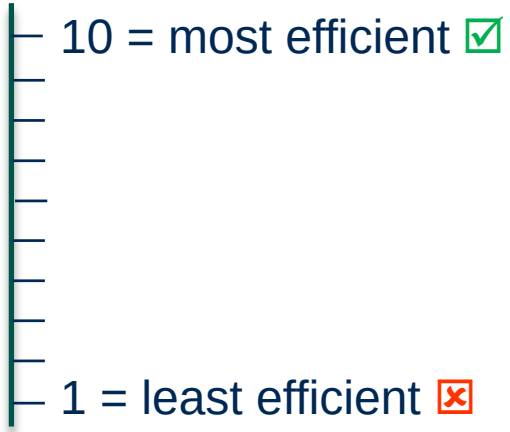



# Step 10: Evaluate Efficiency of Each Test

- Goal: Evaluate the relative efficiency of each test with regard to testing time allotted and testing time used.
- Task 10d: Efficiency (= reliability/mean time spent), i.e., precision per minute spent

| Subtest | Psychometric Efficiency Rating (Precision/Minute) |
|---|---|
| General Science | 8 (0.18) |
| Arithmetic Reasoning | 4 (0.04) |
| Word Knowledge | 10 (0.24) |
| Paragraph Comprehension | 5 (0.07) |
| Mathematics Knowledge | 5 (0.07) |
| Electronics Information | 8 (0.19) |
| Automotive and Shop Information | 8 (0.16) |
| Mechanical Comprehension | 6 (0.10) |
| Assembling Objects | 6 (0.08) |



# Step 11: Synthesize Findings

- Goal: Synthesize findings across all evaluation criteria and tests and summarize the desirability/expendability of each test.
  - Originally intended for Steps 1–10, but the new goal is to synthesize findings over all steps, including the more psychometrically oriented Steps 12–17 (details to follow).
    - Some ratings have been consolidated into a simple spreadsheet (next slide).
  - More slides to follow, summarizing thoughts on methodological approaches to consolidating the evaluation findings into a single rating for each test (Tia Fechter).
- Team: Tia Fechter, Greg Manley, Dan Segall, Mary Pommerich

Next Steps:
- Identify a way to concisely summarize results over all steps.
- Identify a way to aggregate findings and compute an overall rating.

# STEP 11: SYNTHESIZE FINDINGS

## Summary Table for ASVAB Subtest Rating Scales (Steps 4–10)

| Subtests Rating Scale ==> | Total Yearly Cost Rating | Finiteness Rating | Feasibility for AIG Rating | Item Development Success Rate—After Screenings | Item Development Success Rate—After Pool Assembly | Relevancy Rating | Obsolescence Rating | Content Efficiency | Vulnerability Rating | Coachability Rating | Practice Effect Rating | Hardware Effect Rating | Mode Effect Rating | Local Dependence Rating | Familiarity Effect Rating | Psychometric Efficiency* | Average Across All Ratings (1–10 Scale) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| General Science | 3.1 | 6 | 4 | 8.6 | 3.3 | 10 | 7 | 7 | 7 | 10 | 10 | 10 | 10 | 7 | TBD | 8 | 7.4 |
| Arithmetic Reasoning | 6.8 | 10 | 8 | 7.8 | 3.8 | 10 | 10 | 8 | 6 | 10 | 6 | 6 | 10 | 9 | TBD | 4 | 7.7 |
| Word Knowledge | 2.3 | 6 | 7 | 8.4 | 3.9 | 10 | 10 | 10 | 2 | 9 | 10 | 10 | 10 | 9 | TBD | 10 | 7.8 |
| Paragraph Comprehension | 1.0 | 10 | 2 | 7.7 | 2.5 | 10 | 10 | 9 | 7 | 8 | 10 | 10 | 5 | 9 | TBD | 5 | 7.1 |
| Mathematics Knowledge | 6.6 | 8 | 8 | 8.6 | 3.9 | 10 | 10 | 7 | 6 | 10 | 6 | 6 | 8 | 5 | TBD | 5 | 7.2 |
| Electronics Information | 7.5 | 4 | 3 | 8.2 | 3.5 | 6 | 6 | 7 | 8 | 10 | 7 | 10 | 5 | 7 | TBD | 8 | 6.6 |
| Automotive Information | 7.5 | 4 | 1 | 7.8 | 2.6 | 5 | 3 | 9 | 7 | 10 | 10 | 10 | 6 | 7 | TBD | 8 | 6.5 |
| Shop Information | 7.5 | 4 | 1 | 8.2 | 2.5 | 5 | 8 | 9 | 7 | 10 | 10 | 10 | 6 | 7 | TBD | 8 | 6.9 |
| Mechanical Comprehension | 7.1 | 5 | 1 | 8.6 | 3.4 | 8 | 10 | 7 | 8 | 8 | 7 | 5 | 10 | 5 | TBD | 6 | 6.6 |
| Assembling Objects | TBD | 10 | 10 | N/A | N/A | N/A | 10 | 10 | 10 | 5 | 1 | 4 | 9 | 10 | TBD | 6 | 7.7 |

*The AI and SI psychometric efficiency ratings were determined by the psychometric efficiency rating for AS.

# Step 12: Psychometric Impact of Shortening Tests

- Goal: Evaluate the psychometric impact of shortening various tests.
- Task 12a: Review DAC briefings and DAC feedback from prior discussions pertaining to shortening ASVAB in the STP (CEP).
- Task 12b: Evaluate potential impact on CAT-ASVAB test precision.
- Task 12c: Evaluate potential impact on test validity.
- Task 12d: Evaluate potential impact on qualification rates for total group.
- Task 12e: Evaluate potential impact on adverse impact (impact ratios and effect sizes) for demographic groups of interest (as defined in adverse impact analyses).
- Task 12f: Evaluate potential impact on CAT-ASVAB testing time.
- Team: Ping Yin, Mary Pommerich

# Step 12: Psychometric Impact of Shortening Tests

- Goal: Evaluate the psychometric impact of shortening various tests.
- Task 12a: Review DAC briefings and DAC feedback from prior discussions pertaining to shortening ASVAB in the STP (CEP).
- Status:

Completed:

- DAC 1998
  - Discussed shortening ASVAB in the STP (CEP) to accommodate AO
  - Concerns about the increase in time related to adding AO and the impact of shortening tests on reliability and content coverage
- DAC 2000
  - Discussed shortening test lengths to reduce testing time in the schools
  - Recommended that AFQT scores from a shortened battery never be used for enlistment
- DAC 2001
  - Discussed shortening the battery to GS and the AFQT tests in the STP*
  - Concerns about fairness

*Schools were allowed to give just 5 tests, but this practice was not popular and ultimately discontinued.

# Step 12: Psychometric Impact of Shortening Tests

- Goal: Evaluate the psychometric impact of shortening various tests.
- Task 12b: Evaluate potential impact on CAT-ASVAB test precision.
- Status:

Completed:
- i. Used real data to compute latent ability means and SDs for the total group on all subtests (FY18)—see also Step 3b(i)
- ii. Computed measures of precision (reliability) for the current test lengths

Next Steps:
- Simulate CAT-ASVAB data and compute measures of precision (reliability) for selected shortened lengths.

# Step 12: Psychometric Impact of Shortening Tests

- Goal: Evaluate the psychometric impact of shortening various tests.
- Task 12c: Evaluate potential impact on test validity.
- Status:

Completed:
- Summarized content coverage under current test lengths as a baseline for evaluating the impact of shortening test lengths

Next Steps:
- Simulate CAT-ASVAB data and compare content coverage for current test lengths versus shortened test lengths.
- Simulate CAT-ASVAB data and compare intercorrelations between AFQT scores, ASVAB scores, and Service composites for current test lengths versus shortened test lengths.
- Estimate change in validity coefficients for the shortened tests.

# Step 12: Psychometric Impact of Shortening Tests

- Goal: Evaluate the psychometric impact of shortening various tests.
- Task 12d: Evaluate potential impact on qualification rates for total group.
- Status:

Next Steps:

- Simulate CAT-ASVAB data and compare qualification rates for current versus shortened test lengths for:
  - AFQT score
  - Service composites

# Step 12: Psychometric Impact of Shortening Tests

- Goal: Evaluate the psychometric impact of shortening various tests.
- Task 12e: Evaluate potential impact on adverse impact (impact ratios and effect sizes) for demographic groups of interest (as defined in adverse impact analyses).
- Status:

Next Steps:

- Use real data to compute latent ability mean and SD for demographic groups on all subtests.
- Simulate CAT-ASVAB data using N(mean,SD) distributions for demographic groups and compare qualification rates, impact ratios, and effect sizes for current test lengths versus shortened test lengths.

# Step 12: Psychometric Impact of Shortening Tests

- Goal: Evaluate the psychometric impact of shortening various tests.
- Task 12f: Evaluate potential impact on CAT-ASVAB testing time.
- Status:

Next Steps:
- Simulate CAT-ASVAB data and use observed average item latencies to project item and total response times for current versus shortened test lengths.

- More details on Step 12 can be found in the backup slides.

# Step 13: Impact of Using a Math Composite

- Goal: Evaluate psychometric impact of shortening AR and/or MK and computing a composite score (labeled ME) to use in place of AR & MK scores.
- Task 13a: Identify options for shortening AR & MK and computing a composite score (ME).
- Task 13b: Review CAT-ASVAB history for computing AS composite from AI and SI scores.
- Task 13c: Simulate CAT-ASVAB data and compare reliability, qualification rates, and impact ratios for AFQT scores created from VE, AR, & MK versus VE and ME.
- Task 13d: Evaluate impact on Service composites.
- Team: Ping Yin, Mary Pommerich

# Step 13: Impact of Using a Math Composite

- Goal: Evaluate psychometric impact of shortening AR and/or MK and computing a composite score (labeled ME) to use in place of AR & MK scores.
- Task 13a: Identify options for shortening AR & MK and computing a composite score (ME).
- Status:

Next Steps:
- Identify and consider feasible options based on results from Step 12.

# Step 13: Impact of Using a Math Composite

- Goal: Evaluate psychometric impact of shortening AR and/or MK and computing a composite score (labeled ME) to use in place of AR & MK scores.
- Task 13b: Review CAT-ASVAB history for computing AS composite from AI and SI scores.
- Status:

Next Steps:

- Review relevant literature and historical documents on combining AI and SI scores into one composite (AS).

# Step 13: Impact of Using a Math Composite

- Goal: Evaluate psychometric impact of shortening AR and/or MK and computing a composite score (labeled ME) to use in place of AR & MK scores.
- Task 13c: Simulate CAT-ASVAB data and compare reliability, qualification rates, and impact ratios for AFQT scores created from VE, AR, & MK versus VE and ME.
- Status:

Next Steps:
- Compute reliabilities
- Compute qualification rates
- Compute impact ratios

# Step 13: Impact of Using a Math Composite

- Goal: Evaluate psychometric impact of shortening AR and/or MK and computing a composite score (labeled ME) to use in place of AR & MK scores.
- Task 13d: Evaluate impact on Service composites.
- Status:

Next Steps:

- Identify number/degree to which Service composites would be impacted.
- Identify number/degree to which MOSs would be impacted.
- Identify number/degree to which applicants would be impacted.
- Identify alternate composites to replace affected composites and evaluate impact on qualification rates, impact, validity, and classification.

# Step 14: Feasibility of Combining AR/MK into One Test

- Goal: Evaluate the feasibility of combining AR and MK into one test (labeled MA).
- Task 14a: Review AS to AI and SI history to identify potential issues in combining two tests into one.
- Task 14b: Evaluate dimensionality of AR & MK.
- Task 14c: Identify feasible options for combining into a single test.
- Task 14d: Evaluate feasibility/desirability of using multidimensional CAT.
- Team: Furong Gao, Mary Pommerich, Dan Segall, Lihua Yao

# Step 14: Feasibility of Combining AR/MK into One Test

- Goal: Evaluate the feasibility of combining AR and MK into one test (labeled MA).
- Task 14a: Review AS to AI and SI history to identify potential issues in combining two tests into one.
- Status:

Completed:

- Dimensionality studies[1] of the P&P AS data showed statistically significant two factors with low correlation (~0.60).
- The AS pool was split into separate pools for AI and SI items, and the AS (composite) score was derived from unidimensional scoring of AI and SI items separately.

[1]DMDC (2006). *ASVAB Technical Bulletin No. 1: CAT-ASVAB Forms 1 & 2.*

# Step 14: Feasibility of Combining AR/MK into One Test

- Goal: Evaluate the feasibility of combining AR and MK into one test (labeled MA).
- Task 14b: Evaluate dimensionality of AR & MK.
- Status:

Completed:

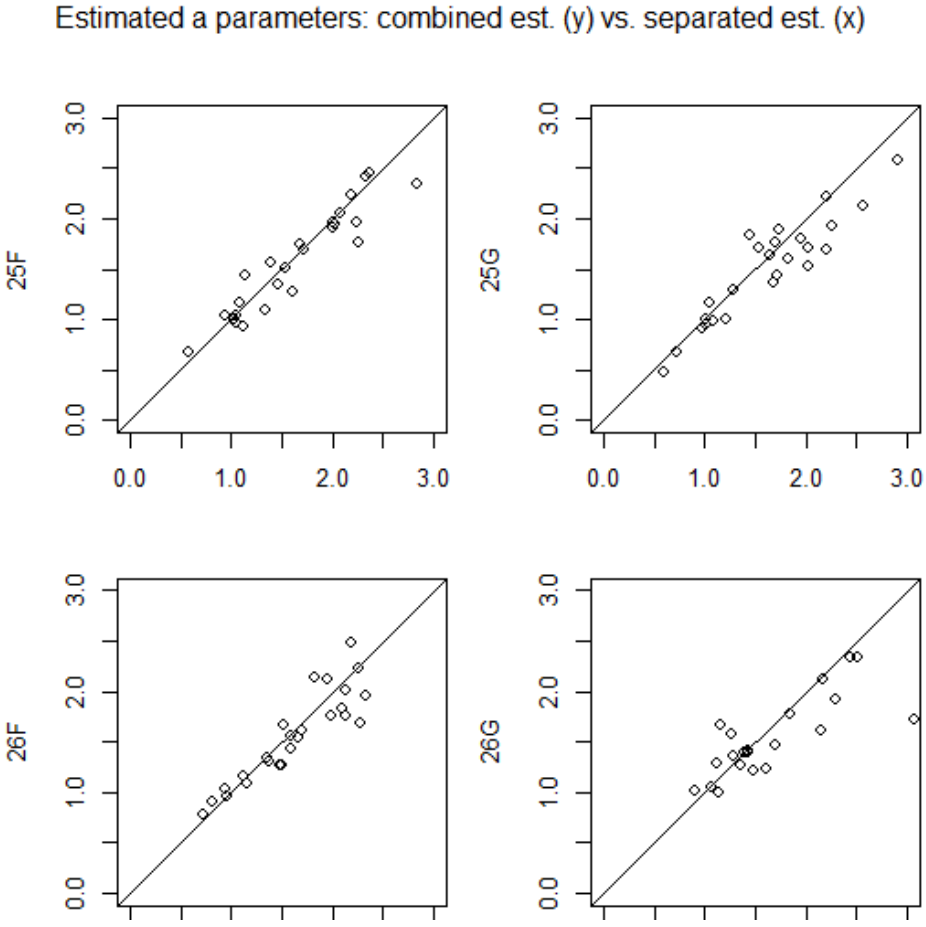
- High correlation confirmed between AR and MK scores.[1]
  - Average correlation value from P&P test scores: ~ 0.72
  - Average correlation value from CAT test scores: ~ 0.74
- Conducted Bilog-MG calibrations on the combined AR/MK data and separate AR/MK data on several P&P forms (25F/G; 26F/G).
  - A unidimensional model fit the combined AR and MK data reasonably well.

[1]DMDC (2008). *ASVAB Technical Bulletin No. 3: CAT-ASVAB Forms 5-9* (Table A.2) and DMDC (2012). *ASVAB Technical Bulletin No. 4: P&P-ASVAB Forms 23-27* (Table F.10).

# Step 14: Feasibility of Combining AR/MK into One Test

- Goal: Evaluate the feasibility of combining AR and MK into one test (labeled MA).
- Task 14b: Evaluate dimensionality of AR & MK.

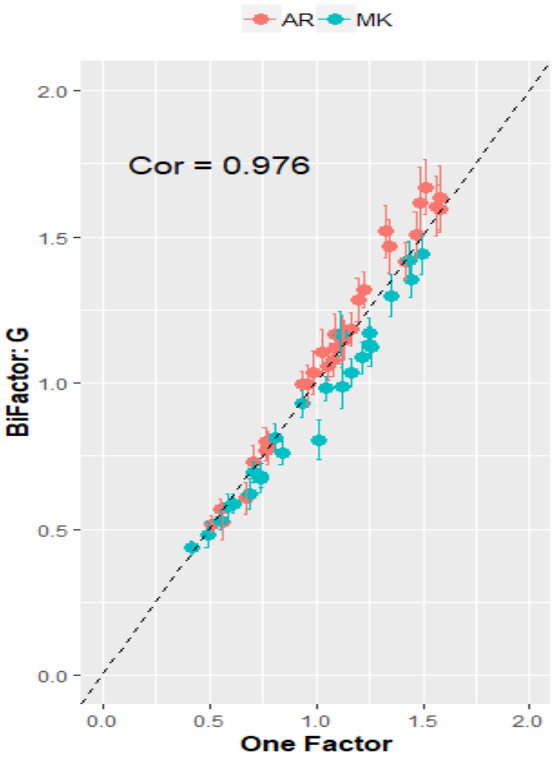
- Correlations of the discrimination parameter estimates from the separate and combined calibrations were all greater than 0.9 except for MK in one P&P form (26G; 0.78).
- Correlations of the difficulty parameter estimates from the separate and combined calibrations were all greater than 0.99.
- Correlations of the guessing parameter estimates were all greater than 0.9.

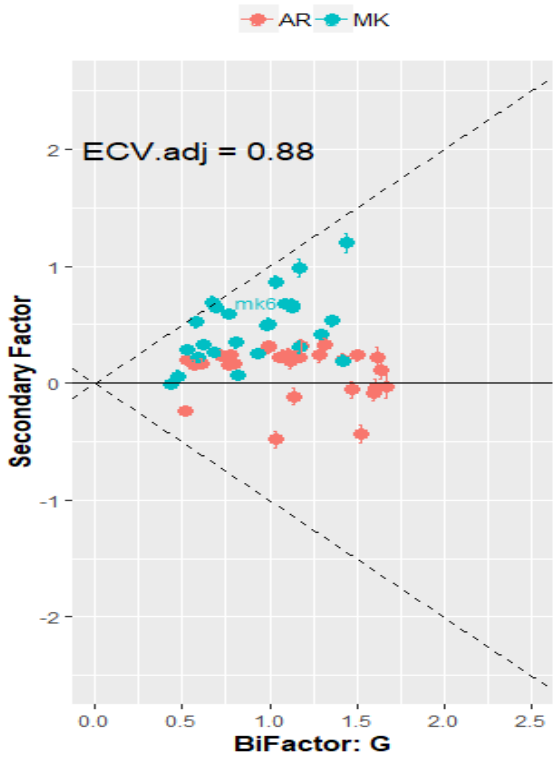


# Step 14: Feasibility of Combining AR/MK into One Test

- Goal: Evaluate the feasibility of combining AR and MK into one test (labeled MA).
- Task 14b: Evaluate dimensionality of AR & MK.

- Confirmatory analyses using iFACT and a bi-factor model also indicated that a unidimensional model would fit the examined data adequately well.
- Correlations from the G-factors of the one-factor and bi-factor models are high.
- Explained common variances (ECV) by the G-factor in the bi-factor model are high (> 0.8).



# Step 14: Feasibility of Combining AR/MK into One Test

- Goal: Evaluate the feasibility of combining AR and MK into one test (labeled MA).
- Task 14c: Identify feasible options for combining into a single test.
- Status:

Next Steps:
- Conduct further investigation on content coverage.
- Examine differential validity of AR and MK.
- Determine item selection algorithm (content balance or split pool).
- Determine an appropriate MA score definition.

- More details on Steps 14a-c can be found in the backup slides.

# Step 14: Feasibility of Combining AR/MK into One Test

- Goal: Evaluate the feasibility of combining AR and MK into one test (labeled MA).
- Task 14d: Evaluate feasibility/desirability of using multidimensional CAT.
- Status:

In Progress:
- A research plan has been established for multidimensional CAT exploration.

Next Steps:
- Develop software for MIRT simulation and conduct simulation.
- Conduct unidimensional simulations for baseline comparisons.
- Check the accuracy of the MIRT software and the models for the MA score.
  - Compare ability recovery and test response times.

- More details on Step 14d plans can be found in the backup slides.

# Step 15: Psychometric Impact of Combining AR & MK

- Goal: Evaluate psychometric impact of combining AR & MK into a single test, labeled MA (assuming prior analyses suggest it is feasible).
- Task 15a: Evaluate potential impact on test validity.
- Task 15b: Evaluate potential impact on CAT-ASVAB test precision.
- Task 15c: Evaluate potential impact on qualification rates for total group.
- Task 15d: Evaluate potential impact on adverse impact (impact ratios and effect sizes) for demographic groups of interest (as defined in adverse impact analyses).
- Task 15e: Evaluate potential impact on CAT-ASVAB testing time.
- Team: Ping Yin, Mary Pommerich

# Step 15: Psychometric Impact of Combining AR & MK

- Goal: Evaluate psychometric impact of combining AR & MK into a single test, labeled MA (assuming prior analyses suggest it is feasible).
- Task 15a: Evaluate potential impact on test validity.
- Status:

Next Steps:

- Simulate CAT-ASVAB data and compare content coverage for AR, MK, and MA for current tests versus selected options.
- Simulate CAT-ASVAB data and compare inter-correlations between AFQT scores, ASVAB scores, MA scores, and Service composites for current test lengths versus selected options.

# Step 15: Psychometric Impact of Combining AR & MK

- Goal: Evaluate psychometric impact of combining AR & MK into a single test, labeled MA (assuming prior analyses suggest it is feasible).
- Task 15b: Evaluate potential impact on CAT-ASVAB test precision.
- Status:

Next Steps:

- Create CAT pools for MA (if feasible).
- Use real data to compute latent ability mean and SD for the total group on all subtests.
- Simulate CAT-ASVAB data and compute for current versus selected options.

# Step 15: Psychometric Impact of Combining AR & MK

- Goal: Evaluate psychometric impact of combining AR & MK into a single test, labeled MA (assuming prior analyses suggest it is feasible).
- Task 15c: Evaluate potential impact on qualification rates for total group.
- Status:

Next Steps:

- Evaluate after completion of Step 12.
- Simulate CAT-ASVAB data and compare qualification rates for current versus selected options for:
  - AFQT score
  - Service composites

# Step 15: Psychometric Impact of Combining AR & MK

- Goal: Evaluate psychometric impact of combining AR & MK into a single test, labeled MA (assuming prior analyses suggest it is feasible).
- Task 15d: Evaluate potential impact on adverse impact (impact ratios and effect sizes) for demographic groups of interest (as defined in adverse impact analyses).
- Status:

Next Steps:

- Use real data to compute latent ability mean and SD for demographic groups on all subtests.
- Simulate CAT-ASVAB data using N(mean,SD) distributions for demographic groups, and compare impact ratios and effect sizes for current test lengths versus selected options.

# Step 15: Psychometric Impact of Combining AR & MK

- Goal: Evaluate psychometric impact of combining AR & MK into a single test, labeled MA (assuming prior analyses suggest it is feasible).
- Task 15e: Evaluate potential impact on CAT-ASVAB testing time.
- Status:

Next Steps:

- Simulate CAT-ASVAB data and use observed average item latencies to project item and total response times for current versus selected options.

# Step 16: Feasibility of Combining EI & Cyber into One Test

- Goal: Evaluate the feasibility of combining EI and Cyber (ICTL) into one test (labeled CE).
- Task 16a: Evaluate content overlap between EI & Cyber.
- Task 16b: Evaluate dimensionality of EI & Cyber.
- Task 16c: Identify feasible options for combining into a single test.
- Task 16d: Evaluate feasibility/desirability of using multidimensional CAT.
- Team: Furong Gao, Mary Pommerich, Dan Segall, Lihua Yao
- Status:

Completed:

- Matched EI and Cyber test data have been obtained for 2018 and 2019.

Next Steps:

- Follow similar approaches used in Step 14 to evaluate feasibility.

# Step 16.5: Feasibility of Combining GS, EI, & MC

- Goal: Evaluate the feasibility of combining the ASVAB technical tests GS, EI, & MC into one test.
- Task 16.5a: Evaluate content overlap between GS, EI, and MC.
- Task 16.5b: Evaluate dimensionality of GS, EI, and MC.
- Task 16.5c: Identify feasible options for combining into a single test.
- Task 16.5d: Evaluate feasibility/desirability of using multidimensional CAT.
- Team: Furong Gao, Mary Pommerich, Dan Segall, Lihua Yao
- Status:

Next Steps:

- Follow similar approaches used in Step 14 to evaluate feasibility.
- Use both P&P and CAT test data whenever possible.

# Step 17: Psychometric Impact of Combining EI & Cyber

- Goal: Evaluate psychometric impact of combining EI & Cyber into a single test, labeled CE (assuming prior analyses suggest it is feasible).
- Task 17a: Create CAT pools for CE.
- Task 17b: Repeat steps outlined earlier for evaluating psychometric impact of combining AR & MK into a single test.
- Team: Ping Yin, Mary Pommerich
- Status:

Next Steps:
- Evaluate after completion of Step 16.

# Step 17.5: Psychometric Impact of Combining GS, EI, & MC

- Goal: Evaluate psychometric impact of combining GS, EI, & MC into a single test (assuming prior analyses suggest it is feasible).
- Task 17.5a: Create CAT pools for the combined test.
- Task 17.5b: Repeat steps outlined earlier for evaluating psychometric impact of combining AR & MK into a single test.
- Team: Ping Yin, Mary Pommerich
- Status:

Next Steps:
- Evaluate after completion of Step 16.5.

# Steps 18–24

- Goal: Evaluate the psychometric impact of dropping existing tests.
- Step 18: Evaluate the psychometric impact of dropping AI.
- Step 19: Evaluate the psychometric impact of dropping SI.
- Step 20: Evaluate the psychometric impact of dropping AO.
- Step 21: Evaluate the psychometric impact of dropping EI.
- Step 22: Evaluate the psychometric impact of dropping MC.
- Step 23: Evaluate the psychometric impact of dropping GS.
- Step 24: Evaluate the psychometric impact of dropping WK.
- Air Force is conducting a related effort that could meet this goal. DPAC will determine the need for additional work upon completion of the Air Force effort.

# Step 25: Synthesize Findings Revisited

- Goal: Synthesize and condense findings/ratings for *all* steps into one rating per test.
  - Some ratings have been consolidated into a simple spreadsheet (Slide 69).
  - Tia Fechter will next summarize thoughts on some methodological approaches to consolidating the evaluation findings into a single rating for each test.

# Consolidating ASVAB Evaluation Findings

# Discussion Topics

- **Need for Synthesis**
- **Three Possible Approaches**
  - Delphi Method
  - Cross-Impact Analysis
  - Utility Analysis
- **Next Steps**

# Need for Synthesis

- Most criteria that subtests are evaluated on have relative degrees of importance compared to one another
- The amount of criteria is too vast to readily make judgments on the relative importance of each ASVAB subtest
- Provides for a more rigorous, research-backed approach for rating the quality (and potential expendability) of the ASVAB subtests

# Delphi Method

- A decision-making approach that engages a panel of experts in providing their opinions on matters of importance for use in determining what could or should be with respect to policy, goal setting, forecasting future outcomes, etc.

- Some interesting facts:
  - Named after the oracle at Delphi (who delivered prophecies)
  - Developed by Norman Dalkey and Olaf Helmer at RAND Corporation in 1950s
  - First uses were for military purposes
    - e.g., determining the number of A-bombs needed to reduce munitions output within various industries
  - Expanded uses
    - Curriculum development
    - Resource utilization
    - Policy determination

# Delphi Method

- The Delphi method is designed to reduce characteristics of group/panel discussions that can hinder concrete decision-making:
  - **Reduces influence of a dominant voice within a group discussion**
    - Antidote: Anonymity
    - Delphi process elicits independent input through a questionnaire format
  - **Inhibits irrelevant or redundant material**
    - Antidote: Controlled Feedback
    - Delphi process is iterative and consists of providing summarized feedback for discussion and consideration before the next round of judgments are made
  - **Mitigates group pressure**
    - Antidote: Calculation of a Statistical Index
    - Delphi process typically makes use of a median to avoid the need for conformity/consensus

# Delphi Method

- Limitations
  - Time-consuming
  - Low response rates will dampen quality of feedback
  - Feedback summary process by investigators may allow investigators to inadvertently impose their own views or biases
  - Knowledge bases of Delphi participants are unevenly distributed
  - Does not account for possible interactions of future events

# Delphi Method

- Example Use: Establishing ASVAB Priorities
  - MAPWG as experts
  - 17 rank-ordered recommendations (e.g., implement CAT at MET sites)
  - Web-based surveys with 5-point Likert scale
  - Elicited ratings on recommendations (intended to improve the relevance of ASVAB) and a set of criteria (timing, costs, benefits)
  - End result was a prioritized list of recommendations

# Delphi Method

- For the ASVAB Evaluation Plan—Possible Synthesis Approach
  - Rate the importance of each scale developed for evaluating ASVAB subtests on various criteria

    e.g., Establish weights for each evaluation criterion based on Delphi process
    - 1.00 = Critically important
    - 0.75 = Considerably important
    - 0.50 = Important
    - 0.25 = Marginally important
    - 0.00 = Not important
  - Use the importance ratings and the scale values assigned by teams for each subtest to determine a single subtest "importance" rating that can be used to rank order subtests (refer to Slide 71)

# Cross-Impact Analysis

- A tool used to evaluate the probability of future events or states; emphasizes the interactions between possible future states, changes, trends, or decisions.

- Some interesting facts:
  - Developed by Theodore Gordon and Olaf Helmer in 1966
  - First use was for Kaiser Aluminum and Chemical Company
    - Game-based approach: “Future”
  - Expanded uses
    - Urban crises
    - Economy simulations
    - Delphi/cross-impact integration
    - Explore policy options

# Cross-Impact Analysis

- Phases
  - **Exploration**: State possible interaction among events
  - **Probabilistic**: Determine how probabilities are elicited
    - Judge events as stand-alone and adjust for cross-impacts post hoc
    - Include possibility of the cross-impacts
  - **Synthesis**: Determine how to collect and summarize judgments
  - **Application**: Collect judgments and evaluate whether adjustments are needed due to non-coherent input (lack of convergence)
    - Implement multiple rounds using either a game-based format or Monte Carlo simulations

# Cross-Impact Analysis

- Limitations
  - Relies on adequacy of pre-determined probabilities and specification of cross-impacts
  - Fatiguing and tedious as number of conditional probability judgments to be made increases (e.g., 10 events crossed = 90 judgments)
  - Accounts for interactions only among pairs of events and not on higher-order effects
- Flexibility
  - In lieu of probabilities,
    - symbolic emphasis/impact may be judged within a matrix, e.g., using coding scheme of ---, --, -, 0, +, ++, +++
    - verbal descriptions may be emphasized, e.g., how does event B impact event A

# Cross-Impact Analysis

- For the ASVAB Evaluation Plan—Possible Synthesis Approach
  - In combination with the Delphi approach, make some determinations about possible interactions
    - If Cyber Test is added, what is the probability of Cyber & EI being combined?
    - If Cyber Test is NOT added, what is the probability of Cyber & EI being combined? [ZERO]
    - If EI is dropped, what is the probability of Cyber Test & EI being combined? [ZERO]
    - If Cyber Test is added and EI is kept, what is the probability that Cyber and EI are NOT combined?
  - Other events
    - Drop subtests (AI, AO, EI, GS, MC, SI)
    - Shorten various subtests
    - Combine AR/MK
    - Add nonverbal measures (MtC, ARt)
    - Add Cyber Test
    - Combine EI/Cyber
    - Add non-cognitive measures (TAPAS, interest inventory)

# Utility Analysis

- A decision-making tool that assigns importance or monetary values to various criteria to evaluate the institutional gain or loss anticipated from various possible courses of action. Decisions are made to maximize benefits while reducing associated costs for those benefits.

- Some interesting facts:
  - a.k.a., Decision Theory, cost–benefit analysis
  - Early contributors include
    - Kelley (1923)
    - Taylor & Russell (1939)
    - Brogden (1946)
    - Naylor & Shine (1965)
    - Cronbach & Gleser (1965)
  - First uses were for business maximization of profit
  - Expanded uses
    - Personnel decisions
    - Military planning
    - Learning experiments

# Utility Analysis

- Quality Indicator: e.g., dollar payoff to the organization for use of a particular decision tool or event

$$\Delta U = \sigma_e r_{ye} \frac{\lambda(y')}{\phi(y')} - \frac{c_y}{\phi(y')}$$

  - $\sigma_e$: standard deviation of the payoff
  - $r_{ye}$: correlation of the predictor and the payoff
  - $\lambda(y')$: ordinate of the normal curve at the cut score on the predictor
  - $\phi(y')$: the upper tail area evaluated at the cut score on the predictor
  - $c_y$: the cost of testing

# Utility Analysis

- Considerations
  - The value of various outcomes needs to be expressed in "equal units of satisfaction," which are additive over many decisions OR must be treated as ordinal
  - Often costly to engage in the required accounting process for the algorithm inputs
  - Some believe the choice of a $ metric leads to a false sense of precision and may be misleading to policy makers
  - The mathematics of Decision Theory are involved and laborious

# Utility Analysis

- For the ASVAB Evaluation Plan—Possible Synthesis Approach
  - Convert all Evaluation Plan metrics to dollar scales, where each indicator of quality is monetized
    - e.g., subtest vulnerabilities, like susceptibility to piracy, can be converted to anticipated costs if piracy were to take place; the scale values could be treated as probabilities, and the cost of piracy could be adjusted based on the likelihood of the piracy to take place
    - e.g., test efficiency can be converted to anticipated cost savings
  - Compare the added utility of each ASVAB configuration in the Evaluation Plan
    - Adding subtests (e.g., Cyber)
    - Dropping subtests (e.g., AI)
    - Combining subtests (e.g., AR/MK)
    - Shortening subtests (e.g., AR/MK)

# Next Steps

- Seek reactions and discussion from MAPWG and DAC regarding proposed approaches.

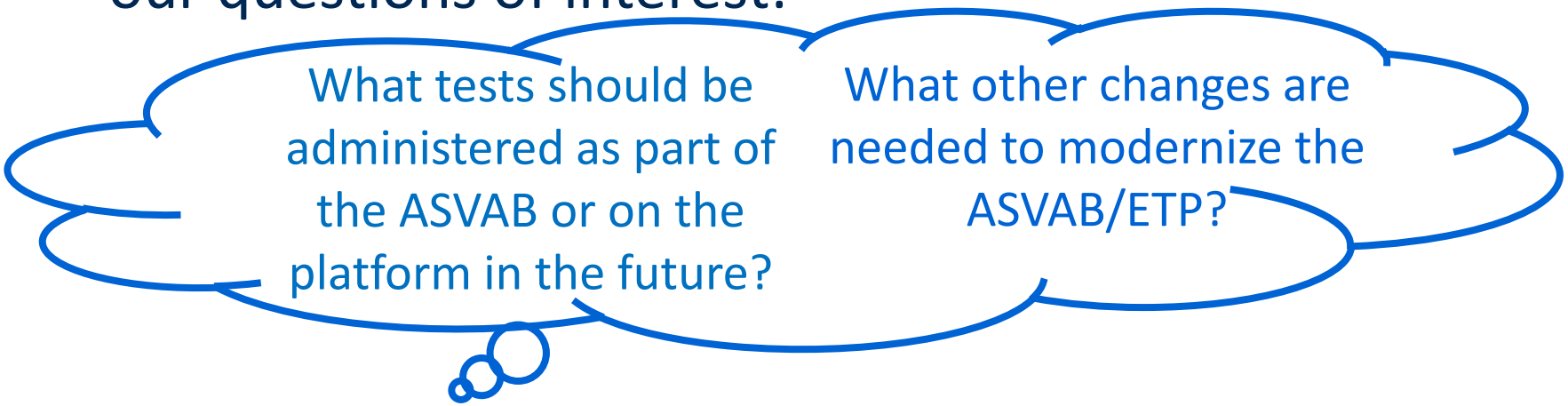
- Trial approach(es) to information synthesis with the goal of arriving at a single importance rating for each ASVAB subtest and special test identified for consideration for inclusion in the ASVAB.

- Report on trial findings.

# Focus Group Effort

# Focus Group Plans

- DPAC is planning a series of focus groups with military testing stakeholders and users.
  - Information gathered will be used to develop a shared vision for *Next Generation Testing*, including our questions of interest:




  - *The hope is to synthesize findings and develop a pathway forward that will converge on a possible solution that will be acceptable to all.*

# MAPWG Focus Group

- An initial focus group was held with the MAPWG.
  - Conducted over a half-day at February's in-person meeting.
  - Sofiya Velgach gave a brief review of the recruiting process and how the ASVAB and special tests are used by the military for selection and classification.
  - Scott Oppler then led a guided group discussion and information gathering.
    - MAPWG members were asked to speak to their perspectives as a MAPWG member, *not* to other stakeholders' perspectives.
    - Participants were told that objective was to collect information—*not* to reach consensus.
    - Some redundancy across questions—could be helpful to consider things from slightly different angles.
    - Participants expressed consent to have the session recorded.

# MAPWG Focus Group

- Key discussion points:

**Topic: Do we need to change the ASVAB/ETP?**

- Identify likes/dislikes about current ASVAB and/or ETP.
- Identify primary reasons for changing ASVAB and/or ETP.
- Identify specific goals to be obtained with a revised ASVAB and/or ETP.

**Topic: What should the ASVAB/ETP predict?**

- Discuss: What should ASVAB/ETP predict?
- Discuss: What outcomes should a revision effort focus on?

**Topic: What would we do if no prior ASVAB?**

- Discuss: If we were going to build the ASVAB from scratch, what would it look like and why?

**Topic: What stakeholders should be involved?**

- Identify stakeholders, users, and other relevant parties we should talk to about how they use the ASVAB/ETP and what their needs are.

# "LIKES" ABOUT THE CURRENT ASVAB/ETP

- Psychometric
  - AFQT predictive of job performance
  - All scales/tests have good measurement precision in addition to predictive validity
  - Has no more adverse impact than other measures of similar constructs
- Content
  - Math/Verbal combination plus technical tests gives ability to do classification (can't just be a cognitive ability test; needs to keep the technical side)
  - Measures both crystalized and fluid intelligence
- Administration
  - Common test among all Services
  - Can be administered in both P&P and CAT
  - PiCAT allows testing prior to MEPS
- Long history and reputation

# "Dislikes" about the current ASVAB/ETP

- Psychometric
  - Adverse impact (although still less than other tests)
  - Not a great balance between population aptitude and what the Services need for some MOSs
  - Score reports aren't easy to understand
- Content
  - Not broad enough—should be expanded to add fluid intelligence and non-cognitive
  - Non-cognitive assessment is part of ETP, but not of ASVAB
    - Would prefer DoD standard, joint-Service assessment
  - Current Cyber items quickly become obsolete

# "Dislikes" about the current ASVAB/ETP

- Administration
  - Administrative time is too long
  - Too many tests required in one session
  - Lack of parallelism in modes of administration for ETP and CEP (CEP requires P&P)
  - PiCAT needs to be proctored to get good results
- Other
  - General perception that ASVAB doesn't change with time (same 10 subtests; no calculator; time since last renorming)
  - The name of the battery ("Vocational" is outdated)
  - Inability for "our side and perspective" to be heard by higher-ups, (e.g., changes to ASVAB would require renorming)
  - Suboptimal communication between research entities

# WHAT IS MISSING FROM THE CURRENT ASVAB/ETP?

- Non-cognitive assessments that measure/evaluate/predict
  - "Propensity to engage in negative behaviors"
  - Honesty/integrity
  - Teamwork
  - Work-environment fit
  - Ability to be trained, accept structure (component of adaptability)
  - "Transfer of training"
  - Verifiable biodata
    - Could be used to assess character (some collected by MEPCOM)
- Measures of
  - Fluid intelligence (need more)
  - Written communication
  - Situational judgment
  - Problem-solving
  - Emotional intelligence
  - Cyber *aptitude* (as opposed to knowledge)
  - Psychomotor
  - Cognitive multi-tasking

# WHAT IS EXPENDABLE IN THE CURRENT ASVAB/ETP?

- Less use for General Science (GS) and Electronics Information (EI)—for some Services
- Assembling Objects (AO)—there is a perception that only one Service uses it
- Auto Information–because of adverse impact and content

# Other improvements that could be made

- More unproctored testing
- Automated Item Generation (AIG) to reduce item development time
- Combining subtests to reduce test administration time

# PRIMARY REASONS FOR CHANGING THE ASVAB/ETP

- Psychometric
  - Increase incremental validity over AFQT
  - Increase classification efficiency—want tests that have differential validity across job types
  - Increase diversity
  - Make more resistant to compromise
- Content
  - Training needs/nature of work may have changed
  - Assess the "whole person"
  - Better identify people who fit the current military culture
- Administration
  - Reduce testing time
  - Take advantage of technological advances (both IT and measurement)

# PRIMARY REASONS FOR CHANGING THE ASVAB/ETP

- Perceptions
  - Fix perceptions that “the ASVAB doesn’t change with time” and “we aren’t measuring the right things”
  - Everybody else is changing (and lack of change is hurting us from a leadership perspective)
  - Increase face validity (for Congress/Generals)
- Workforce
  - Increase numbers of eligible applicants
  - Change policies that meet the requirements for the total workforce

# REASONS FOR *NOT* CHANGING ("BARRIERS")

- Costs (research)
  - Classification composites would need to be re-established/validated
  - Re-norming would be required
  - Volumes of research would be negated
- Costs (non-research)
  - IT changes would be required (e.g., flow of data and required changes in each database to accommodate different programming languages)
  - Documentation would need to be changed
- Other
  - Emotional impact of change; fear; ripple effect creates more work
  - Current ASVAB/ETP still meets the primary needs
  - Lack of consensus on what changes to make
  - Maintaining comparability with CEP

# Mitigations to barriers for change

- Utility analyses to demonstrate ROI
- Frequent communication with stakeholders/users
- Stakeholder/user buy-in
- Eliminate/remove pervasive misrepresentations

# Goals to be obtained with a revised ASVAB/ETP

- Psychometric
  - Increase prediction, differential prediction, and classification efficiency
  - Maintain/improve reliability
  - Maintain a stable scale score
  - Reduce adverse impact
  - Increase representativeness and inclusion (to reflect the population)
  - Reduce probability of compromise
  - Make scoring easier to understand (CEP)
    - Simplify score reporting (CEP and maybe ETP)
- Content
  - Increase breadth of coverage

# Goals to be obtained with a revised ASVAB/ETP

- Administration
  - Reduce testing time
  - Increase flexibility to administer at home
- Perceptions
  - Improve face validity
  - Improve perceptions of the test; correct misperceptions
    - e.g., CAT-ASVAB is too hard, P&P ASVAB is easier, etc.
- Other
  - Improve cost-efficiency

# What would a brand new ASVAB/ETP look like?

- New name
  - "Vocational" is an outdated term
- Content*
  - Include g in "core" tests (used by all for enlistment eligibility—like those in current AFQT)
  - Better balance between crystallized and fluid intelligence
  - Include spatial/psychomotor
  - Include non-cognitive measures (some potentially in core)
    - Include interests, but not in core
  - Include Cyber as a technical test in ASVAB (not as a special test)
  - Include physical/occupational assessment
  - See everything on "What is Missing" list

*Note: A job analysis might be useful to identify constructs to include.

# What would a brand new ASVAB/ETP look like?

- Structure
  - Approximately 90 minutes of testing (on average)
  - Reduced number of "mandatory" tests
    - e.g., include only core tests in ASVAB; treat all others (including technical) as special tests
- Administrative/Delivery Protocols*
  - Adaptive tests administered on computers
  - New item types that take advantage of technology
    - Expand beyond multiple-choice items

*Note: The question of providing accommodations falls outside of the scope of the MAPWG.

# What would a brand new ASVAB/ETP look like?

- Collect Ancillary Information
  - Biodata
  - Parental educational attainment
  - Proxies for socio-economic status (SES)
  - Language capabilities
  - English as a second language (ESL) status
  - Presidential Physical Fitness Test performance

# WHAT SHOULD THE ASVAB/ETP PREDICT?

- Training success
  - Recycles, attrition from training, course grades, setbacks, failures
- On-the-job performance
- Attrition
- First-term re-enlistment
- Promotions and promotion rates
- Talent management (long-term success)
- Person-job/environment fit
- Organizational citizenship behaviors
- Physical fitness
- Attitude
- Leadership potential
- Commitment
- Teamwork
- Adaptability/agility
  - Integrate new capabilities; adapt warfighting approaches; ability to work within existing doctrine to accomplish mission; change business practices

# What outcomes should a revision focus on?

- Highest priority (3-way tie)
  - Completion of training
    - No setbacks, pass on first attempt
  - Job performance
    - Ability; person-job fit
  - Attrition
    - Non-EAS (end of active service) attrition
- Next highest priorities
  - Increase classification efficiency
  - Reduce adverse impact

# Stakeholders Identified for Targeted Focus Groups

**Military/DoD Realm**

- Accession Policy/MEPCOM
- Service policy-makers
- Military trainers
- Classifiers
- Recruiters
- National Guard Bureau
- Recruiting operations commands
- Functional/community/occupational managers

**Educational Realm**

- Educational Service Specialists
- Department of Education
- State Boards of Education
- Career counseling organizations
- High school counselors
- High school and community college teachers

**Wish List**

- Congress

**Examinee Realm**

- Applicants/recruits
- Students and influencers

Focus group protocols will be tailored to the target audience.

# Focus Group Future Steps

- Identify SMEs for focus groups.
- Develop protocols for focus groups.
- Schedule and conduct focus groups.
  - Will be conducted virtually, which could impact duration.
- Compile results across focus groups.
- Form and convene an ASVAB Stakeholder Advisory Committee (ASAC) to help guide decision-making about *Next Generation Testing.*
  - Will comprise SMEs from the various stakeholder groups.

# Next Steps for Next Generation Testing

# Next Steps

- DPAC will continue in their efforts to complete the ASVAB evaluation steps outlined in this briefing.
- DPAC will begin efforts to evaluate special tests of interest in a similar fashion.
- DPAC will continue efforts to establish a methodological approach to rating the quality/expendability of the ASVAB tests.
- DPAC will continue efforts to develop a shared vision for *Next Generation Testing*.